Highlights

A Comprehensive Evaluation Framework for Synthetic Trip Data Generation in Public Transport

Yuanyuan Wu, Zhenlin Qin, Zhenliang Ma

- A comprehensive `RPU` (*Representativeness-Privacy-Utility*) framework is proposed to evaluate synthetic trip data in public transport.
- The framework assesses synthetic data across three dimensions and three hierarchical levels (record, group, population).
- Twelve representative generation methods are benchmarked, including statistical, deep generative, and privacy-enhanced approaches.
- Results show that synthetic data do not inherently guarantee privacy, and no 'one-size-fits-all' model exists.
- CTGAN achieves the most balanced trade-off between representativeness, privacy, and utility, making it suitable for practical applications.

A Comprehensive Evaluation Framework for Synthetic Trip Data Generation in Public Transport

Yuanyuan Wu[a], Zhenlin Qin[a], Zhenliang Ma[b,*]

[a] *Division of Transport Planning, Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, Sweden*

[b] *Transport Planning Division and Digital Futures, KTH Royal Institute of Technology, Stockholm, Sweden*

Abstract

Synthetic data offers a promising solution to the privacy and accessibility challenges of using smart card data in public transport research. Despite rapid progress in generative modeling, there is limited attention to comprehensive evaluation, leaving unclear how reliable, safe, and useful synthetic data truly are. Existing evaluations remain fragmented, typically limited to population-level representativeness or record-level privacy, without considering group-level variations or task-specific utility. To address this gap, we propose a *Representativeness-Privacy-Utility* (`RPU`) framework that systematically evaluates synthetic trip data across three complementary dimensions and three hierarchical levels (record, group, population). The framework integrates a consistent set of metrics to quantify similarity, disclosure risk, and practical usefulness, enabling transparent and balanced assessment of synthetic data quality. We apply the framework to benchmark twelve representative generation methods, spanning conventional statistical models, deep generative networks, and privacy-enhanced variants. Results show that synthetic data do not inherently guarantee privacy and there is no "one-size-fits-all" model, the trade-off between privacy and representativeness/utility is obvious. Conditional Tabular generative adversarial network (CTGAN) provide the most balanced trade-off and is suggested for practical applications. The `RPU` framework provides a systematic and reproducible basis for researchers and practitioners to compare synthetic data generation techniques and select appropriate methods in public transport applications.

Keywords: Synthetic trip data, Evaluation model, Data privacy, Data representativeness, Data utility, Generative models

1. Introduction

Smart card data have become essential for understanding travel behavior, demand patterns, and policy impacts in public transit systems (Pelletier et al., 2011). These data provide rich, passively collected records of passenger movements, enabling large-scale and deep empirical analysis. However, growing concerns over data privacy have led public transport agencies and data owners hesitate to share data or apply anonymization techniques such as masking,

*Corresponding author at: Brinellvägen 23, SE-100 44 Stockholm, Sweden. Tel.:+46 8 790 48 14

Email addresses: `yuanwu@kth.se` (Yuanyuan Wu), `zhenlinq@kth.se` (Zhenlin Qin), `zhema@kth.se` (Zhenliang Ma)

noise injection, or data swapping. While these methods reduce disclosure risk, they often significantly degrade the utility of the data for the downstream analysis and decision-making (Kieu et al., 2023).

In response, synthetic data generation using using machine learning and deep generative models (DGMs) has gained growing attention (Choi et al., 2025). Techniques such as generative adversarial networks (GANs) (Goodfellow et al., 2014) and variational autoencoders (VAEs) (Huang et al., 2019) strive to reproduce the statistical properties of real data, while also balancing the protection of sensitive or identifiable information.

Despite growing interest in synthetic data, most research has focused on developing generative models, with relatively limited attention to evaluating whether the generated data are both useful and privacy-preserving. A systematic and robust evaluation framework remains remains underdeveloped. Such a framework is critical for both researchers and practitioners to assess whether synthetic data are fit for purpose in real-world transport applications. In its absence, model validation, comparison, and practical adoption remain ad hoc and inconsistent. Evaluating synthetic data requires answering three fundamental questions:

1. Does the data accurately reflect the real-world distributions?
2. Does it sufficiently protect sensitive or identifiable information?
3. Is it still useful for downstream analysis and decision-making?

These questions correspond to three core dimensions of synthetic data quality that we refer to as: *representativeness*, *privacy*, and *utility*. However, current evaluation practices in the literature remain limited in several important respects. First, assessments are heavily skewed toward representativeness, typically measured through distributional similarity at the population level (e.g., marginal or joint distributions) (Kapp et al., 2023). Individual and group level representativeness, which are critical for capturing heterogeneous travel behaviors and subgroup dynamics, are often overlooked. Second, privacy evaluations are narrowly focused, usually limited to duplicate checks or record-level membership inference attacks (MIA) (Smolak et al., 2020). Few studies consider group-level privacy, such as leakage of sensitive signals from identifiable subpopulations, or population-level privacy, which concerns the inference of aggregate patterns or behavioral trends. Third, utility is often conflated with representativeness and assessed using similarity metrics. While some degree of similarity is necessary for utility, relying solely on distributional similarity provides an incomplete picture of whether the synthetic data are fit for downstream use. The true utility of synthetic data requires dedicated evaluation, particularly for tasks such as monitoring, planning, or policy analysis.

Moreover, the conflation of metrics across evaluation dimensions is problematic. For example, high distributional similarity is sometimes interpreted as both good representativeness/utility and low privacy risk (Berke et al., 2022). It is an assumption that does not always hold. Such mixed or overlapping use of metrics can lead to misleading conclusions and hinders rigorous comparison between models or informed selection for deployment.

This paper addresses these limitations by proposing a multi-dimensional, multi-level evaluation framework for synthetic public transport data, based on dedicated indicators (Figure 1). We apply this framework to benchmark a diverse set of statistical, deep generative, and privacy-enhanced models, using real-world AFC trip data. Empirical insights into generative model performance and the trade-offs involved in producing synthetic trip data for

public transport applications are derived.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 develops the evaluation framework. Section 4 describes the generative models for benchmarking. Section 5 demonstrates its use for synthetic trip data generation in public transport. Section 6 summarizes the key findings and future research.

2. Related work

2.1. Background

Synthetic data refers to artificially created records that mimic patterns and structures observed in real-world datasets (Drechsler, 2011). These datasets are typically produced by fitting generative models to the original data in order to capture its key statistical characteristics, either the joint probability $p(x, y)$ or the marginal distribution $p(x)$. Once trained, such models can be used to generate new data instances, such as travel trajectories, that resemble the original distribution while allowing flexible scaling (Goncalves et al., 2020). A wide range of generative modeling techniques have been explored in the literature, ranging from classical probabilistic methods to modern deep learning-based approaches, representing as Deep Generative Models (DGMs) (Choi et al., 2025).

Classical models, such as Gaussian mixture models (Reynolds, 2015), Gaussian copulas (Nelsen, 2006), and Bayesian networks (BNs) (Koller and Friedman, 2009), fit explicit probabilistic distributions directly to real data. While these models are interpretable and effective in many applications, they typically rely on simplifying assumptions (e.g., Gaussianity, conditional independence) and may struggle to capture complex, high-dimensional, or non-linear patterns in real-world data such as mobility behaviors. Nevertheless, they provide strong and transparent baselines for evaluating the performance of DGMs. Moreover, their underlying principles continue to influence modern architectures. For example, GMMs are used in Conditional Tabular GANs (CTGAN) to sample training batches and mitigate mode collapse (Xu et al., 2019); copula-based normalization has been proposed to improve model transferability across domains (Jutras-Dubé et al., 2024); and BNs have been leveraged to derive topological orderings that guide the autoregressive generation process in large language models (LLMs) for population synthesis (Lim et al., 2025).

Unlike classical models, DGMs leverage neural network architectures to flexibly approximate non-linear dependencies and latent structure in the data without relying on strong parametric assumptions. Two widely used families of DGMs for generating mobility data are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Variational Autoencoders (VAEs) (Kingma et al., 2013). GANs consist of a generator and a discriminator trained in opposition: the generator aims to produce realistic synthetic data, while the discriminator attempts to distinguish between real and synthetic samples. Variants such as CTGAN adapt this framework for tabular data by conditioning on specific feature values and using training tricks to reduce mode collapse (Xu et al., 2019). Extensions like PATE-GAN incorporate differential privacy by training the discriminator on an ensemble of teacher models with noise injection to protect individual data contributions (Jordon et al., 2018). VAEs are probabilistic models that encode data into a latent space and reconstruct samples by decoding from this space (Kingma et al., 2013). They optimize a variational lower bound on the data likelihood, balancing reconstruction accuracy with latent space regularization. Compared to

GANs, VAEs are generally more stable to train and offer a structured, interpretable latent space (Doersch, 2016). However, they may produce less sharp or over-regularized samples if not carefully tuned (Lucic et al., 2018), especially for categorical or structured data (Bao et al., 2022).

In recent work, more expressive architectures have been proposed. For instance, diffusion models iteratively denoise random noise toward realistic samples through learned reverse processes, achieving state-of-the-art performance in image synthesis and increasingly adapted for tabular data generation (Croitoru et al., 2023; Yang et al., 2023). Normalizing flows transform a simple distribution into a complex one through a sequence of invertible mappings, allowing exact likelihood estimation (Papamakarios et al., 2021). Large Language Models (LLMs) have also been explored as autoregressive generators for structured records, leveraging token-level conditioning and fine-tuning strategies (Patel et al., 2024; Lim et al., 2025).

DGMs have been widely applied across various urban mobility data generation tasks, including population synthesis (Badu-Marfo et al., 2022; Kim and Bansal, 2023), synthetic traffic data (Wang et al., 2020; Li et al., 2022), origin-destination (OD) matrices generation (Rong et al., 2023a,b), and vehicle or human trajectory modeling (Huang et al., 2019; Fontana et al., 2023), among others. For broader overviews, readers are referred to the reviews by Kapp et al. (2023) and Choi et al. (2025).

2.2. DGMs for smart card data synthesis

In the context of smart card data synthesis, GANs are the most commonly used models for smart card data synthesis, while VAEs often serve as baseline benchmarks (Badu-Marfo et al., 2022). For example, Kim et al. (2022) used a conditional GAN to impute qualitative attributes of trip chains from smart card data, showing improved fidelity, diversity, and creativity compared to a conditional VAE and other baselines, as measured by Jensen-Shannon Distance (JSD) and confusion matrices. Privacy concern was not included in their evaluation.

Focusing more directly on synthetic trip generation, Kieu et al. (2023) utilized a tabular GAN (CTGAN) to synthesize bus smart card trip records from South-East Queensland's transit system. Their study compared the performance of CTGAN with that of a Bayesian network using over one million real trip records. Representativeness was evaluated by examining how well the synthetic data replicated key statistical distributions. While both models captured general travel patterns, CTGAN tended to overestimate peak-hour travel and rare trip types, whereas the Bayesian network provided a closer fit for attributes such as travel time distributions. For privacy evaluation, the authors conducted a duplicate detection check by identifying any synthetic records that exactly matched real trips. Although this method provides a basic indication of overfitting or memorization, it falls short of offering a comprehensive or formal privacy assessment.

To address privacy more rigorously, Badu-Marfo et al. (2020) developed a differentially private (DP) GAN for generating daily travel activity diaries. Their model incorporated differential privacy via DP-SGD during training, thereby ensuring that individual-level information in the training data was protected. In addition to evaluating the representativeness through marginal distributions and pairwise correlations, they employed membership inference attacks (Shokri et al., 2017) to quantify the level of privacy risk. While their study focused on household travel surveys (HTS) data, the approach is adaptable to smart card data synthesis.

HTS datasets provide rich behavioral and socioeconomic information, making them valuable for modeling urban mobility. However, their utility is constrained by several well-known limitations: they are expensive and time-consuming to collect, suffer from recall bias, are infrequently updated, and are increasingly restricted by privacy concerns. To address these challenges, synthesizing smart card (SC) data using generative models offers a promising alternative, as SC data provide large-scale, high-resolution observations collected passively over time (Uğurel et al., 2024). Recent studies have explored data fusion strategies that combine the complementary strengths of SC data. For example, Xu et al. (2024) proposed an GAN-based framework to infer the age attribute extracted from SC data and examine the relationship between built environment and age using features from Points of Interest data. Vo et al. (2025) proposed a cluster-based data fusion method that leverages SC data to enhance the spatial granularity of HTS datasets, which are limited by low sampling rates. Similarly, Lee et al. (2025) introduced a collaborative GAN-based framework to fuse HTS and SC data for generating detailed activity schedules. While these works evaluate the representativeness and utility of the synthesized data, they do not include formal privacy assessments.

2.3. Synthetic data evaluation

Privacy is commonly stated in the literature as the main motivation for applying DGMs to mobility data synthesis. However, in practice, many studies focus primarily on demonstrating the feasibility or novelty of their generative approaches, with evaluation metrics tailored to specific modeling goals. To demonstrate the effectiveness of proposed DGMs in reproducing diverse population, Kim and Bansal (2023) used standardized root mean square error (SRMSE) for marginal and bivariate distributions, and confusion metrics (precision, recall, F1) to indicate feasibility and diversity. Zhang et al. (2023) assessed modality-aware trajectory generation via Jensen-Shannon Divergence (JSD) on modality distributions and transitions, while Kieu et al. (2023) used the Wasserstein distance to compare spatial distributions. In benchmarking CTGAN, TVAE, and Gaussian Copulas for carsharing trip synthesis, Albrecht et al. (2024) evaluated fidelity through statistical tests (Kolmogorov-Smirnov, Chi-squared) and graphical comparisons of marginal distributions and pairwise correlations. These metrics help quantify how well the synthetic data preserve the structural characteristics of the real data, especially for tabular data.

Privacy concerns are frequently addressed in studies on trajectory generation. For example, Kulkarni et al. (2018) benchmarked RNNs, GANs, and Gaussian copulas in generating synthetic mobility trajectories, evaluating the privacy–utility tradeoff using membership inference attacks (MIA). Similarly, Fontana et al. (2023) introduced a privacy-aware framework that combines a GAN for trajectory generation with a deep learning-based user identification model. Privacy was evaluated through identification accuracy, following the data uniqueness concept proposed by Sweeney (2002). More recently, Zhang et al. (2025) proposed a diffusion model incorporating collaborative noise priors for urban mobility synthesis. They evaluated the model using marginal distribution alignment, privacy protection via uniqueness testing and MIA, and utility through downstream prediction performance.

Overall, most existing studies assess the quality of synthetic data using only one or two dimensions, typically focusing on either *statistical similarity* (e.g., distributions, pairwise distances) or *utility* in downstream tasks (e.g., prediction accuracy). Common fidelity measures such as Jensen-Shannon Divergence (JSD) and Wasserstein distance are widely applied

to evaluate how well the synthetic data mirrors real data at distribution/population-level. Frameworks like SDV (Patki et al., 2016) support benchmarking across statistical models and GAN-based generators, with an emphasis on utility and general similarity metrics. In contrast, privacy is often overlooked, and few studies comprehensively evaluate synthetic data across representativeness, privacy, and utility, especially at multiple levels. Recent efforts have begun to adopt more multi-dimensional evaluation strategies. For example, Platzer and Reutterer (2021) introduced a holdout-based fidelity and privacy assessment for mixed-type tabular data using individual distance-to-nearest-record as a proxy for privacy risk. Eigenschink et al. (2023) identified five key criteria for evaluating synthetic data quality: representativeness, novelty, realism, diversity, and coherence. Similarly, Lautrup et al. (2025) proposed SynthEval, an open-source framework integrating fidelity and privacy metrics across both categorical and numerical features for unified benchmarking. However, despite these advances, evaluations are often limited to the population/distribution level, neglecting how synthetic data performs at finer granularities such as individual records or within specific user groups.

Table 1 provides an overview of representative studies on *synthetic mobility data* evaluation, checking their focus on representativeness, utility, and privacy. As shown, most works prioritize representativeness at population-level and downstream utility, while systematic privacy assessment remains limited. This imbalance highlights the need for a more comprehensive and structured evaluation across both dimensions and levels.

Table 1: Summary of related work applying DGMs for synthetic mobility data

Study	Model Type	Benchmark(s)	Data Type	Use Case	Evaluations		
					R*	P*	U*
Kulkarni et al. (2018)	GAN	Copulas, RNNs	GPS	generate trajectory	✓	✓	×
Borysov et al. (2019)	VAE	Gibbs sampler, BNs	HTS	synthesize population	✓	×	×
Badu-Marfo et al. (2020)	DP-GAN		HTS + GPS	synthesize travel diary	✓	✓	×
Kim et al. (2022)	cGAN	Gibbs sampler, BN, VAE	SC + HTS	impute sociodemographic attributes and trip purposes	✓	×	×
Kieu et al. (2023)	CTGAN	BN	SC	synthesize trips	✓	✓	×
Kim and Bansal (2023)	WGAN	BN, VAE	HTS	synthesize population	✓	×	×
Zhang et al. (2023)	GAN	other GANs	GPS	generate trajectory	✓	×	×
Fontana et al. (2023)	privacy-aware GAN	N.A.	GPS	generate trajectory	✓	✓	✓
Albrecht et al. (2024)	Multiple	CTGAN, TVAE, Copula	carsharing data	synthetic tabular data	✓	×	×
Xu et al. (2024)	GAN	Random Forest, BN, VAE	SC + PoI	infer socioeconomic characteristics of passengers	✓	×	×
Lee et al. (2025)	GAN		SC + HTS	generate activity schedules	✓	×	✓
Zhang et al. (2025)	Diffusion	GANs, Diffusion	GPS	generate trajectory	✓	✓	✓
This study	Multiple	GANs, VAEs, etc.	SC	synthetic trips	✓	✓	✓

*R: Representativeness; P: Privacy; U: Utility

3. RPU evaluation framework

To comprehensively assess the quality of synthetic data, we propose a structured evaluation framework organized along three key dimensions: *representativeness*, *privacy*, and *utility*, denoted as RPU framework (Figure 1, Table 2). Each dimension is assessed through multiple aspects, reflecting distinct perspectives of data quality.

3.1. Representativeness and indicators

The representativeness dimension, denoted $\mathcal{R}$, measures how well the synthetic data replicates key characteristics of the real data. This includes *population-level* similarity in global distributions and structural relationships ($\mathcal{R}_p$), *group-level* fidelity in capturing subgroup-specific patterns ($\mathcal{R}_g$), and *record-level* plausibility and internal coherence ($\mathcal{R}_r$).

Table 2: Definition and indicators of RPU evaluation framework

Representativeness		
Aspects	**Definition**	**Indicators**
Record-level plausibility	The extent to which each synthetic data record exhibits plausible, coherent, and realistic behavior when considered as a stand-alone entity.	- Visual inspection - Rule-based logic validation - Expert judgment
Group-level heterogeneity	The degree to which the synthetic data preserves the distinctive behavioral patterns and statistical properties of subpopulations defined by shared characteristics.	- KLD, JSD, EMD across groups
Population-level similarity	The extent to which the overall statistical structure of the synthetic data matches that of the original dataset, including marginal distributions, joint feature dependencies, and spatial-temporal patterns.	- KLD, JSD, EMD
Privacy		
Aspects	**Definition**	**Indicators**
Record-level disclosure	The risk that an individual synthetic record reveals the membership or attributes of a particular real individual.	- Visual inspection - MIA - Attribute inference
Group-level exposure	The risk of reproducing rare or vulnerable groups, revealing group-level patterns that could be exploited, or amplifying the exposure of marginalized populations.	- K-NN distance ratio across groups
Population-level leakage	The risk that synthetic data may leak sensitive structural, behavioral, and societal patterns of the underlying population.	- K-NN distance ratio
Utility		
Aspects	**Definition**	**Indicators**
Monitoring	Use synthetic data to track trends, detect anomalies, and support operational dashboards.	- Time-series similarity - Peak location accuracy - Change-point detection
Predictive modeling	Use synthetic data to train or evaluate models that predict future or unknown outcomes.	- TSTR - Accuracy, AUC, F1 - Feature importance correlation
Policy analysis	Use synthetic data to simulate or evaluate effects of interventions or interruptions.	- Expert evaluation

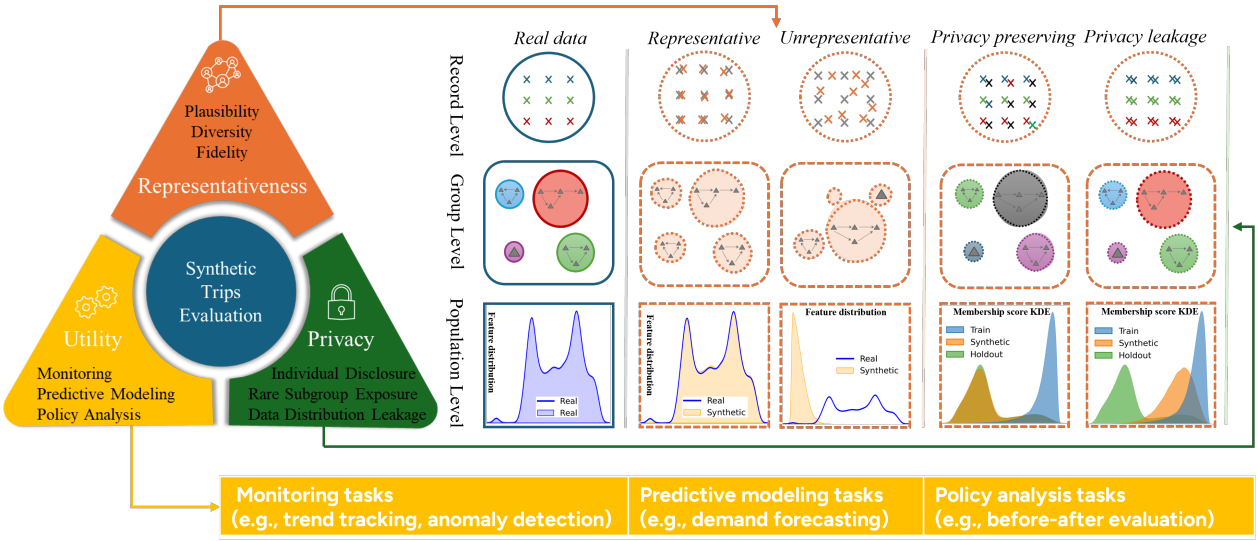

Figure 1: Representativeness, privacy, utility evaluation framework for synthetic trips.

A fundamental requirement for synthetic data is that its distribution $P(x')$ should closely approximate the distribution of the original data $D(x)$. This similarity is typically quantified using statistical divergence or distance measures. One widely used metric is the *Kullback-Leibler divergence (KLD)*, defined as:

$$\mathrm{KL}(D(x) \,\|\, P(x')) = \begin{cases} \sum_x D(x) \log \frac{D(x)}{P(x')} & \text{(discrete case)} \\ \int D(x) \log \frac{D(x)}{P(x')}\, dx & \text{(continuous case)} \end{cases} \tag{1}$$

KL divergence measures the information loss when $P(x')$ is used to approximate $D(x)$. It is non-negative and equals zero if and only if the two distributions are identical almost everywhere. A **smaller** KL value indicates higher similarity between the two distributions. KLD is asymmetry and not bounded. The *Jensen-Shannon divergence (JSD)* is often used as a symmetric and smoothed alternative:

$$\begin{aligned} \mathrm{JSD}(D(x) \,\|\, P(x')) &= \frac{1}{2}\mathrm{KL}(D(x) \,\|\, M(x)) + \frac{1}{2}\mathrm{KL}(P(x') \,\|\, M(x)) \\ \text{where } M(x) &= \frac{1}{2}(D(x) + P(x')) \end{aligned} \tag{2}$$

JSD is a symmetric and bounded divergence measure. When using log base 2, which is common in information-theoretic contexts, its values lie within the range $[0, 1]$, facilitating interpretability and comparability across models. Another widely adopted metric, especially in the evaluation of generative models, is the *Wasserstein distance*, which computes the cost of moving one distribution to match another, with smaller distances indicating closer distributions, is commonly used to evaluate the spatial distribution similarity:

$$W(D, P) = \inf_{\gamma \in \Pi(D,P)} \mathbb{E}_{(x,x') \sim \gamma} \left[\|x - x'\| \right] \tag{3}$$

where $\Pi(D, P)$ denotes the set of all joint distributions (couplings) with marginals $D(x)$ and $P(x')$. Wasserstein distance, is also known as the *Earth Mover's Distance (EMD)*, provides

meaningful gradients even when the distributions do not overlap, making it particularly suitable for training and evaluating DGMs.

While many existing studies focus primarily on population-level evaluation, assessing *how well the overall statistical structure of synthetic data aligns with the original data* using metrics such as KLD, JSD, and EMD. Such aggregate measures may overlook important variations at finer granularities. Specifically, high distributional similarity does not guarantee that synthetic records are individually plausible or that subpopulation characteristics are faithfully reproduced (Eigenschink et al., 2023).

To provide a more comprehensive assessment, our framework incorporates two additional levels of evaluation. The *group-level* component assesses whether the synthetic data preserves distinct behavioral patterns across meaningful subsets of the data, defined by features such as peak vs. off-peak hours, weekdays vs. weekends, or user-defined criteria. This is measured using conditional KLD, JSD, and EMD. The *record-level* is defined as the degree to which synthetic data retain natural plausibility, even if statistical properties are well-matched. Assessment at this level is performed through visual inspection, rule-based validation, and expert judgment, focusing on the behavioral plausibility of each synthetic trip and the internal consistency of its attributes, such as valid transfer sequences and temporally coherent travel timelines.

3.2. Privacy and indicators

We evaluate privacy also at three levels: the *record level* ($\mathcal{P}_r$), such as identity and attribute disclosure; the *group level* ($\mathcal{P}_g$), including rare subgroup exposure; and the *population level* ($\mathcal{P}_p$), including structural or behavioral pattern leakage.

To assess record-level privacy risk, we adopt a customized membership inference attack (MIA) tailored to our synthetic data evaluation setting (Algorithm 1), which estimate the probability $p_{mia}(x)$ that a given record $x \in D_{real}$ was part of the training data. Specifically, we train a binary classifier to distinguish between real training records (labeled as members) and holdout records (non-members) based on selected feature vectors. The attack model is then applied to synthetic data, where all records are presumed non-members. We report the mean predicted membership probability across synthetic records, which reflects the average confidence of the classifier that synthetic samples resemble training data. This approach provides a practical and interpretable signal of overfitting or memorization by the generative model. Additionally, we visualize the distributions of membership scores for training, holdout, and synthetic samples using kernel density estimation (KDE) plots, enabling a qualitative comparison of risk exposure across datasets.

Algorithm 1 Membership Inference Attack via Random Forest Classifier

Require: Real training data D_{train}, real holdout data D_{holdout}, synthetic data D_{syn}, feature list $\mathcal{F}$

function PREPAREMIADATA($D_{\text{train}}, D_{\text{holdout}}, \mathcal{F}$)
    Assign label 1 to D_{train}, label 0 to D_{holdout}
    Combine both datasets into D_{combined}
    $X \leftarrow D_{\text{combined}}[\mathcal{F}]$, $y \leftarrow D_{\text{combined}}[\text{label}]$
    return X, y
end function
function TRAINMIACLASSIFIER(X, y)
    Split X, y into training and test sets with stratified sampling
    Train Random Forest classifier $\mathcal{C}$ on training set
    Predict membership scores y_{pred} on test set
    Compute AUC: AUC $\leftarrow$ roc_auc($y_{\text{test}}, y_{\text{pred}}$)
    return $\mathcal{C}$, AUC
end function
function EVALUATEONSYNTHETIC($\mathcal{C}, D_{\text{syn}}, \mathcal{F}$)
    $X_{\text{syn}} \leftarrow D_{\text{syn}}[\mathcal{F}]$
    Compute membership scores: $s \leftarrow \mathcal{C}.predict\_proba(X_{\text{syn}})[:, 1]$
    return Mean confidence: $\mu_s \leftarrow \text{mean}(s)$
end function
function PLOTMIASCOREDISTRIBUTION($\mathcal{C}, D_{\text{train}}, D_{\text{holdout}}, D_{\text{syn}}, \mathcal{F}$)
    Compute scores $s_{\text{train}}, s_{\text{holdout}}, s_{\text{syn}}$ using $\mathcal{C}$ on each dataset
    Combine scores with labels into one data structure
    Plot KDE of score distributions for each type
end function

Divergence-based measures (e.g. KLD, JSD, and EMD) are commonly used in the literature to assess the population or sub-group level leakage, since overly close distribution matches imply privacy risks. However, because these measures are already employed to evaluate representativeness in our framework, we instead evaluate population- and group-level privacy risks using a *k-nearest neighbor* (k-NN) distance-based analysis (Algorithm 2). At the population level, we compute the average distance from each synthetic record to its (k) nearest real records, and compare this to the average distance between each real record and its ($k + 1$) nearest real neighbors (excluding the self-match). The ratio between these two quantities serves as an interpretable metric of memorization or overfitting risk. A lower ratio indicates that synthetic records are more tightly clustered around real data than expected, suggesting potential privacy leakage.

To extend this analysis to the group level, we stratify the data by a grouping variable ($\mathcal{F}$) and compute the k-NN distance ratio within each group. This group-wise analysis captures variations in privacy risk across different segments of the data distribution.

Algorithm 2 k-NN Distance-Based Privacy Evaluation

Require: Real data D_{real}, Synthetic data D_{syn}, Feature list $\mathcal{F}$, Optional group column g, Number of neighbors k, Thresholds $(\theta_{\text{low}}, \theta_{\text{high}})$

function EvaluatePopulationLevelPrivacy($D_{\text{real}}, D_{\text{syn}}, \mathcal{F}, k$)
  Fit k-NN model on $D_{\text{real}}[\mathcal{F}]$
  Compute distances from each synthetic record to k nearest real records: $d_{\text{syn}\to\text{real}}$
  Compute mean over all distances: $\bar{d}_{\text{syn}\to\text{real}}$
  Fit k+1-NN model on $D_{\text{real}}[\mathcal{F}]$
  Compute distances from each real record to its $k+1$ nearest real neighbors
  Exclude self-distance and compute mean: $\bar{d}_{\text{real}\to\text{real}}$
  Compute distance ratio: $\rho = \frac{\bar{d}_{\text{syn}\to\text{real}}}{\bar{d}_{\text{real}\to\text{real}}}$
  return ρ
end function
function EvaluateGroupLevelPrivacy($D_{\text{real}}, D_{\text{syn}}, \mathcal{F}, g, k$)
  Extract unique group values: $G \leftarrow \text{unique}(D_{\text{real}}[g])$
  Initialize list $\mathcal{R} \leftarrow []$ ▷ To store group-wise ratios
  for all $v \in G$ **do**
    $D^{(v)}_{\text{real}} \leftarrow$ real records with $g = v$
    $D^{(v)}_{\text{synth}} \leftarrow$ synthetic records with $g = v$
    if $|D^{(v)}_{\text{real}}| > k$ **and** $|D^{(v)}_{\text{syn}}| > 0$ **then**
      Compute $\rho^{(v)} \leftarrow$ EvaluatePopulationLevelPrivacy($D^{(v)}_{\text{real}}, D^{(v)}_{\text{syn}}, \mathcal{F}, k$)
      Append $\rho^{(v)}$ to $\mathcal{R}$
    else
      Skip group v due to insufficient data
    end if
  end for
  return Group-level mean ratio: $\bar{\rho}_{\text{group}} = \text{mean}(\mathcal{R})$
end function

3.3. Utility and indicators

The utility dimension, denoted as $\mathcal{U}$, evaluates how well synthetic data can support meaningful downstream analytical tasks compared to the real data. Therefore, the utility can be assessed across representative categories, including *monitoring* (e.g., trend detection and system surveillance), *predictive modeling*(e.g., training and validating machine learning models) and *policy analysis* (e.g., *'before-after'* evaluations, scenario simulations). In practice, utility should be assessed in a task-specific manner. For example, clustering utility can be evaluated by applying K-Means separately to real and synthetic data and comparing the resulting centroids using Euclidean distance. For predictive modeling, a common strategy is *Train on Synthetic, Test on Real* (TSTR), which evaluates how well models trained on synthetic data generalize to real observations. The resulting performance is then compared to a baseline established by *Train on Real, Test on Real* (TRTR), providing a reference for the predictive utility gap between synthetic and real data.

4. Generative models for benchmarking

4.1. Statistical models

Classical examples of generative models include Gaussian Mixture Models (GMMs), Gaussian copulas, and Bayesian networks (BN). GMMs assume that data is generated from a mixture of several Gaussian distributions and estimate parameters through expectation-maximization (Reynolds, 2015). Real data is used to estimate the parameters of the distributions and then the synthetic data can be produced by sampling a component according to the learned mixture weights and drawing a sample from the corresponding Gaussian distribution.

Gaussian Copulas separate the modeling of marginal distributions from their dependency structure. In the first step, each variable's marginal distribution is estimated and transformed into a standard normal space via the probability integral transform. Then, a multivariate Gaussian distribution is fit to capture the dependencies between transformed variables (Nelsen, 2006). Synthetic data is generated by sampling from the fitted multivariate normal distribution and then inverting the transformation back to the original data scale using the estimated marginals.

Bayesian Networks (BNs) represent the joint distribution of a set of variables using a directed acyclic graph (DAG), where each node corresponds to a variable and edges encode conditional dependencies. The structure of the graph is learned from data, and conditional probability tables (CPTs) are estimated for each node given its parents (Koller and Friedman, 2009). Synthetic data is generated by ancestral sampling: variables are sampled in topological order based on the graph, using the learned CPTs.

4.2. Generative adversarial networks

Generative Adversarial Networks (GANs) are a class of deep generative models that learn to produce synthetic data by framing the generation process as a two-player minimax game between a generator and a discriminator (Goodfellow et al., 2014). Let $G(z;\theta_G)$ be a generator that maps random noise $z \sim p_z(z)$ to synthetic samples, and $D(x;\theta_D)$ be a discriminator that outputs the probability that input x is a real trip rather than a synthetic one. The objective function is formulated as:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

During training, the discriminator D tries to distinguish real trips $x \sim p_{\text{data}}$ from generated ones $G(z)$, while the generator G attempts to fool the discriminator by producing trips $x' = G(z)$ that resemble real ones. It is proven in Goodfellow et al. (2014) that the dual problem in Eq. 4 is equivalent to minimizing the JSD between the generated distribution and the real distribution. However, the training suffers from issues such as vanishing gradients when minimizing $\log(1 - D(G(z)))$, which can hinder effective learning. To address this, the Wasserstein GAN (WGAN) (Arjovsky et al., 2017) was proposed replacing the Jensen–Shannon divergence with the Wasserstein-1 distance to provide smoother gradients and more stable training. This was further improved by adding a gradient penalty in WGAN-GP (Gulrajani et al., 2017). In this study, WGAN-GP is adopted and evaluated due to its improved convergence properties and robustness.

Conditional GANs (cGANs) extend GAN framework by conditioning both G and D on additional side information c:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x \mid c)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z \mid c)))] \tag{5}$$

In the transit context, c could represent features such as the intended origin station, travel time interval, or passenger category. Conditioning has the potential to improve the model's ability to generate context-aware synthetic trips, increasing their behavioral plausibility and control. Conditional Tabular GAN (CTGAN) (Xu et al., 2019), which explicitly incorporates conditioning during training to better handle mixed data types and imbalanced distributions, falls into this category and is benchmarked in this study.

4.3. Variational autoencoders

Variational Autoencoders (VAEs) are probabilistic generative models that learn a latent representation of data and generate new samples by decoding from this latent space (Kingma et al., 2013).

A VAE consists of two neural networks: an encoder $q_\phi(z \mid x)$, which approximates the posterior distribution over latent variables z given input data x, and a decoder $p_\theta(x \mid z)$, which reconstructs the input data from the latent space. The VAE is trained to maximize the evidence lower bound (ELBO) on the marginal log-likelihood of the data:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \text{KL}(q_\phi(z \mid x) \| p(z)) \tag{6}$$

where $\text{KL}(\cdot\|\cdot)$ denotes the KLD between the approximate posterior and the prior $p(z)$, typically taken as a standard normal distribution $\mathcal{N}(0, I)$. The first term encourages accurate reconstruction of the input data, while the second term regularizes the latent space to conform to the prior distribution, promoting smoothness and generalization.

A **conditional VAE (cVAE)** extends VAE framework by conditioning both encoder and decoder on observed features c, such as temporal context or user profiles:

$$\log p_\theta(x \mid c) \geq \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x \mid z, c)] - \text{KL}(q_\phi(z \mid x, c) \| p(z \mid c)) \tag{7}$$

For public transit applications, the input of VAEs and cVAEs may consist of structured features such as origin, destination, departure time, route characteristics, or user type. The decoder learns to generate synthetic trips that reflect these characteristics while capturing latent variability in the underlying behavior. Both VAE (Kingma et al., 2013) and cVAE (Xu et al., 2019) are benchmarked in this study.

4.4. Diffusion models

Diffusion models are a class of generative models that synthesize data by reversing a gradual noising process, learning to transform pure noise into structured samples through a sequence of denoising steps (Ho et al., 2020). These models have recently achieved state-of-the-art results in image and time-series generation and are now being explored in mobility domains for their ability to capture fine-grained temporal dynamics and distributional diversity.

The diffusion process consists of two stages: a forward process (diffusion), where noise is incrementally added to the data, and a reverse process (generation), where the model learns

to progressively denoise and recover the original data. Formally, the forward process defines a Markov chain $q(x_t \mid x_{t-1})$ that gradually adds Gaussian noise to a data sample x_0 over T time steps:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I) \tag{8}$$

where $\beta_t \in (0, 1)$ controls the noise schedule.

The reverse process is parameterized by a neural network $\epsilon_\theta(x_t, t)$, trained to predict and remove the added noise. The generative model approximates the reverse transition as:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{9}$$

Training involves minimizing a variational bound, which in practice reduces to a simplified denoising score matching objective:

$$\mathcal{L}_{\text{simple}}(\theta) = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, t)\right\|^2\right] \tag{10}$$

where $\bar{\alpha}_t$ is the cumulative product of $(1-\beta_t)$, and $\epsilon \sim \mathcal{N}(0, I)$.

In the context of synthetic public transit trips, diffusion models can be applied to learn distributions over trip features or sequences, such as travel times, origins, or trajectory embeddings. Their iterative nature allows for modeling multimodal distributions and uncertainty in generation, which is valuable in replicating diverse mobility patterns. However, diffusion models present challenges in practice (Peng et al., 2024). They they are computationally intensive due to the large number of sampling steps, and they require careful architectural design to incorporate domain-specific constraints. To this end, a multilayer perceptron (MLP)-based denoiser is customized and benchmarked in this study.

4.5. Normalizing flow

Normalizing Flows (NFs) are a class of likelihood-based generative models that transform a simple base distribution into a complex target distribution through a series of *invertible* and *differentiable* mappings.

Let $\mathbf{x} \in \mathbb{R}^d$ denote a real-valued data vector (e.g., a trip record), and $\mathbf{z} \in \mathbb{R}^d$ be a latent variable sampled from a known prior $p_{\mathbf{Z}}(\mathbf{z})$, typically a standard multivariate normal distribution. A normalizing flow defines a sequence of K invertible functions f_k such that:

$$\mathbf{x} = f_K \circ f_{K-1} \circ \cdots \circ f_1(\mathbf{z}) = f(\mathbf{z}), \quad \mathbf{z} = f^{-1}(\mathbf{x}) \tag{11}$$

The log-likelihood of a data point $\mathbf{x}$ is given by the change-of-variables formula:

$$\log p_{\mathbf{X}}(\mathbf{x}) = \log p_{\mathbf{Z}}(f^{-1}(\mathbf{x})) + \sum_{k=1}^{K} \log \left|\det\left(\frac{\partial f_k^{-1}}{\partial \mathbf{h}_k}\right)\right|, \quad \text{where } \mathbf{h}_k = f_k^{-1} \circ \cdots \circ f_1^{-1}(\mathbf{x}) \tag{12}$$

the Jacobian determinant quantifies the volume change introduced by each transformation.

In our implementation, we adopt a RealNVP-style flow (Dinh et al., 2016) using affine coupling layers. Each coupling layer splits the input $\mathbf{x} = [\mathbf{x}_A, \mathbf{x}_B]$ and applies the following transformation:

$$\mathbf{y}_A = \mathbf{x}_A,$$
$$\mathbf{y}_B = \mathbf{x}_B \odot \exp(s(\mathbf{x}_A)) + t(\mathbf{x}_A),$$

where $s(\cdot)$ and $t(\cdot)$ are neural networks producing scaling and translation parameters, and $\odot$ denotes element-wise multiplication. The triangular Jacobian structure enables efficient training via exact likelihood computation.

To synthesize trips, the NF model is trained by maximizing the log-likelihood on real trip vectors, consisting of normalized continuous features and embedded or one-hot categorical features. Synthetic data is generated by sampling $\mathbf{z} \sim \mathcal{N}(0, I)$ and applying the forward transformation $\mathbf{x} = f(\mathbf{z})$. Outputs are postprocessed to decode categorical variables (e.g., origin, destination) from softmax outputs or discrete embeddings.

4.6. LLM as generators

Large Language Models (LLMs), such as GPT and BERT variants, have demonstrated impressive generative capabilities across domains including text, code, and structured data (Raiaan et al., 2024). More recently, LLMs have been explored for synthetic data generation in tabular and sequential settings, including human mobility modeling, by treating structured records as tokenized sequences and learning the joint distribution through language modeling objectives (Borisov et al., 2022; Lim et al., 2025).

In the context of public transit, synthetic trips can be formulated as sequences of discrete tokens, where each token represents a feature or event, such as origin station, departure time, travel mode, or transfer point. A synthetic trip T can be expressed as a sequence:

$$T = \{w_1, w_2, \ldots, w_n\} \tag{13}$$

where each token $w_i \in \mathcal{V}$ is drawn from a vocabulary $\mathcal{V}$ of possible trip elements (e.g., station codes, time bins, user types).

LLMs are trained using an autoregressive objective that models the probability of the next token given the preceding ones:

$$P(T) = \prod_{i=1}^{n} P(w_i \mid w_1, \ldots, w_{i-1}) \tag{14}$$

Fine-tuning or prompt engineering can incorporate contextual conditions such as user type, day of the week, or policy scenario. In such cases, the model can be conditioned on a prefix c, modifying the generation as:

$$P(T \mid c) = \prod_{i=1}^{n} P(w_i \mid c, w_1, \ldots, w_{i-1}) \tag{15}$$

This generative approach enables flexible sampling and customization, as well as natural integration of heterogeneous data. For example, trip chains, passenger attributes, and even travel policies can be embedded within token sequences or prompt templates. In this study, we customize and benchmark the generative LLM proposed by Zhao et al. (2025).

4.7. Privacy-enhanced models

Differential Privacy (DP) (Dwork et al., 2006) provides a formal mathematical framework for quantifying the privacy guarantees of algorithms that release information about datasets. A randomized algorithm $\mathcal{A}$ satisfies ε-differential privacy if, for any two neighboring datasets D and D' differing in at most one record, and for any measurable subset of outputs $S \subseteq \text{Range}(\mathcal{A})$, the following condition holds:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(D') \in S] \tag{16}$$

Here, $\varepsilon > 0$ is the *privacy budget*, with smaller values implying stronger privacy protection. This ensures that the inclusion or exclusion of a single individual's data has a limited influence on the algorithm's output.

A more relaxed version, known as (ε, δ)-differential privacy, allows for a small probability δ that the privacy guarantee is violated:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(D') \in S] + \delta \tag{17}$$

This relaxation is often necessary in deep learning-based mechanisms (Abadi et al., 2016) such as the Gaussian mechanism or teacher-student aggregation schemes. In our study, we consider both definitions depending on the generative model used.

To benchmark privacy-preserving enhanced synthetic data generation, we include two representative models that incorporate differential privacy guarantees using distinct methodological approaches: statistical Priv-BN (Zhang et al., 2017) and neural PATE-GAN (Jordon et al., 2018). **Priv-BN** (PrivBayes) constructs a differentially private Bayesian Network to approximate the joint distribution of tabular data. It iteratively learns a set of low-dimensional conditional distributions by building a directed acyclic graph over the attributes. To achieve ε-differential privacy, Laplace noise is added to the computation of marginal and conditional probabilities during both the structure learning and parameter estimation stages. The final synthetic dataset is generated by sampling from the noisy Bayesian network. This approach offers interpretability and strong privacy guarantees, though it may struggle with capturing complex, high-dimensional dependencies. **PATE-GAN** integrates the Private Aggregation of Teacher Ensembles (PATE) framework into a GAN architecture to achieve (ε, δ)-differential privacy. The dataset is split into disjoint subsets, each used to train a teacher classifier (discriminator). Noisy majority voting across these teachers provides labels used to train a student model, ensuring differential privacy via the Gaussian mechanism. This student model is then used to guide the discriminator in adversarial training. The generator learns to produce data that can fool the discriminator while remaining differentially private through the PATE aggregation. PATE-GAN is particularly suited to high-dimensional data but introduces trade-offs between utility and the amount of noise required to protect privacy.

5. Case study and results

5.1. Data description

The Mass Transit Railway (MTR) in Hong Kong accommodates an average of 5 million passenger trips per day, with the vast majority paid through the Octopus Card, a contactless smart card embedded within the automatic fare collection (AFC) system. Key trip details

such as anonymized card IDs, card type, tap-in and tap-out times/stations, and fare deductions are recorded in the transaction data. The ubiquity, comprehensive coverage, and high spatiotemporal resolution of such data have made it a valuable resource for advancing transit modeling (Nassir et al., 2019; Mo et al., 2021), operational intelligence (Halvorsen et al., 2020), passenger behavior analysis (Ma et al., 2020), and public policy evaluation (Wu et al., 2025). However, access to this data is restricted due to privacy concerns, limiting its potential for open research and collaboration.

In this study, we use Octopus card data as the real reference dataset and evaluate the performance of various generative models in producing synthetic trip data. We aim to assess whether synthetic data can adequately represent real data, support similar downstream applications, and offer reduced privacy risks for data sharing. The real dataset used in this study consists of anonymized Octopus card transactions collected over a two-week period from May 14 to May 27, 2018 (Figure 2). This period was deliberately selected to capture a diverse range of travel patterns influenced by regular weekdays, weekends, and a public holiday (May 22, Buddha's birthday). The selected period allows us to observe several anticipated variations in travel behavior, including typical weekday commuting patterns, regular weekend travel, holiday-related behaviors, and differences between normal and holiday-affected workdays or weekends.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
14	15	16	17	18	19	20
21	22	23	24	25	26	27
May 21: Potential bridge day before public holiday			May 22: Public holiday (Buddha's Birthday)			

Figure 2: Calendar breakdown of selected data period (May 14–27, 2018).

From the raw dataset, we extracted three disjoint subsets for model training, testing, and validation. Each record in the extracted data includes the following features: anonymized passenger ID, trip origin, destination, start time, end time, and the corresponding day of the week as conditional information. The training set consists of 9,000 passengers with a total of 140,725 trips, the testing set includes 3,011 passengers with 47,283 trips, and the validation set comprises 2,985 passengers with 46,395 trips. The sampling procedure was designed to preserve the original distributions of trip count per passenger, as well as the temporal characteristics of start time and end time. The distributional consistency is illustrated in Figure 3.

5.2. Calculation and normalization of evaluation scores

To ensure fairness in comparison, all models are trained on the same input data and are required to generate an equal number of synthetic trips, denoted as (T_{syn}). No post-processing is applied to the generated data, allowing for a direct assessment of raw model performance.

We first assess record-level representativeness by applying logical consistency checks to each generated trip. Specifically, we examine whether the generated origin-destination (OD) pairs exist in the original data, whether the start time precedes the end time, and whether time values fall within the valid range (0–1440 minutes, as all times are encoded in minutes

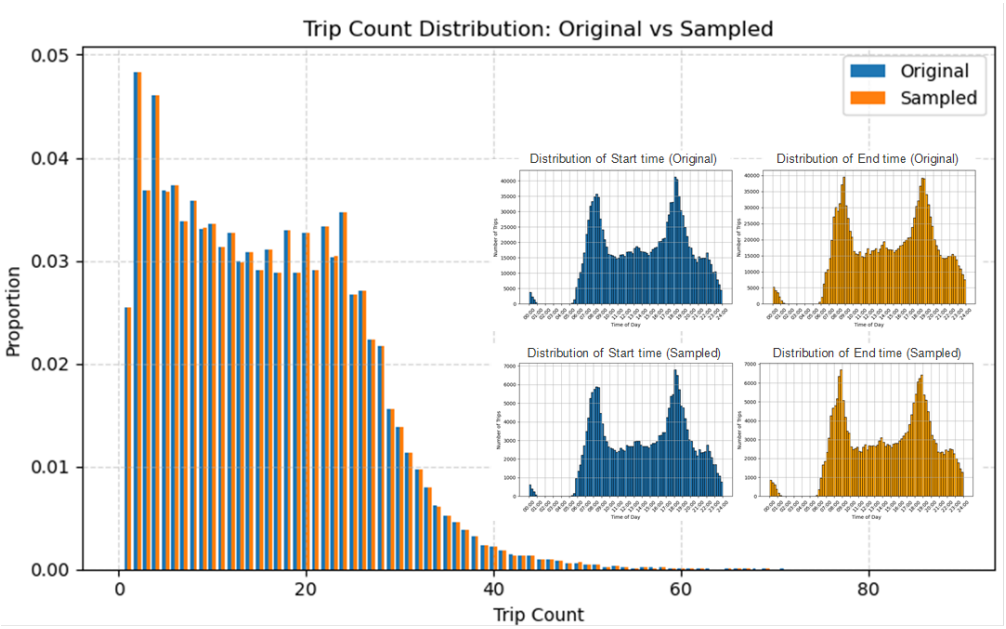

Figure 3: The sampled datasets preserve the original distribution of trip counts, start time and end time.

past midnight). Records that fail any of these checks are counted as invalid and denoted by T_{syn_f}. The record-level representativeness score, $\mathcal{R}_r$, is then defined as the proportion of logically consistent records:

$$\mathcal{R}_r = 1 - \frac{T_{syn_f}}{T_{syn}}$$

This metric reflects the model's ability to generate plausible individual records consistent with real-world constraints.

Population-level representativeness is evaluated using three statistical distance measures: Kullback–Leibler divergence (KLD), Jensen–Shannon divergence (JSD), and Earth Mover's Distance (EMD) as introduced in subsection 3.1, denoted as $\mathcal{R}_{p_k}$, $\mathcal{R}_{p_j}$, and $\mathcal{R}_{p_e}$, respectively. Similarly, group-level representativeness is assessed by comparing distributions conditioned on the day of the week, using the same three metrics: $\mathcal{R}_{g_k}$, $\mathcal{R}_{g_j}$, and $\mathcal{R}_{g_e}$. To ensure comparability across models, we apply *min-max normalization* to each metric across all models. The final group- and population-level representativeness scores are then computed as the average of the normalized metrics:

$$\mathcal{R}_g = \text{Average}(\mathcal{R}_{g_k}, \mathcal{R}_{g_j}, \mathcal{R}_{g_e}), \quad \mathcal{R}_p = \text{Average}(\mathcal{R}_{p_k}, \mathcal{R}_{p_j}, \mathcal{R}_{p_e})$$

The overall representativeness score for a given model m is computed as the average of its record-level, group-level, and population-level scores:

$$\mathcal{R}^m = \text{Average}(\mathcal{R}_r^m, \mathcal{R}_g^m, \mathcal{R}_p^m)$$

Privacy evaluation is performed using membership inference attacks (MIA) and k-nearest neighbor (k-NN) analysis, as detailed in Algorithm 1 and 2. Higher membership probability indicates greater privacy risk; thus, the mean MIA probability of each model is normalized to

p_{mia}, and the record-level privacy score is defined as $\mathcal{P}_r = (1 - p_{mia})$. In the k-NN analysis, a larger distance ratio corresponds to lower privacy risk. The resulting distance ratios are normalized to obtain the population-level $\mathcal{P}_p$ and group-level $\mathcal{P}_g$ privacy scores. The overall privacy score for model m is computed as the average of its record-level, group-level, and population-level scores:

$$\mathcal{P}^m = \text{Average}(\mathcal{P}_r^m, \mathcal{P}_g^m, \mathcal{P}_p^m)$$

The utility evaluation is task-specific. In this study, we assess two downstream tasks: clustering and prediction. To assess clustering utility, K-Means clustering is applied separately to real and synthetic datasets, and the average minimum distance between their centroids is computed. A smaller distance indicates higher clustering utility. For predictive utility, a *Train on Synthetic, Test on Real* (TSTR) evaluation is conducted by training a Gradient Boosting Regressor with k-fold cross-validation on both datasets and evaluating on real test folds. The difference in predictive accuracy, measured by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), between TSTR and a *Train on Real, Test on Real* (TRTR) baseline quantifies predictive utility. Smaller differences indicate better preservation of modeling utility. All utility indicators are first normalized across models: d_c (clustering distance), d_{mae}, and d_{rmse} (prediction differences). The scores are defined as: $\mathcal{U}_c^m = 1 - d_c$ (clustering score), $\mathcal{U}_p^m = \text{Average}(1 - d_{mae}, 1 - d_{rmse})$ (prediction score), the overall utility score is computed as:

$$\mathcal{U}^m = \text{Average}(\mathcal{U}_c^m, \mathcal{U}_p^m).$$

It is important to note that all scores shown in the following figures are **min-max normalized** across models; therefore, lower values indicate comparatively weaker performance in a given dimension relative to other models, rather than poor absolute performance.

5.3. Benchmarking results

We present the benchmarking results of the proposed evaluation framework applied to a wide range of generative models as introduced in section 4. To begin, we compare all benchmarked methods across the three evaluation dimensions.

5.3.1. Evaluation across all methods

Figure 4 presents the normalized evaluation metrics across all methods, revealing that no single model achieves uniformly strong performance across representativeness, privacy, and utility. The substantial variation across record-, group-, and population-level metrics highlights that relying on a single evaluation level may obscure important performance differences, reinforcing the importance of multi-level assessment.

Figure 5 presents the overall score for each dimension, along with the average across representativeness, privacy, and utility for each method. Deep generative models (DGMs) do not surpass conventional statistical methods in overall performance, although they exhibit a slight advantage in privacy protection. A more promising direction is to improve DGM performance by leveraging strengths from conventional approaches. Incorporating GMM-based sampling to guide training and mitigate mode collapse, CTGAN is a good example to show the potential of this strategy. As a result, it achieves the highest overall performance, while ranking just behind Priv-BN and PATE-GAN in privacy protection. But these two privacy-enhanced models failed in the evaluation of representativeness and utility.

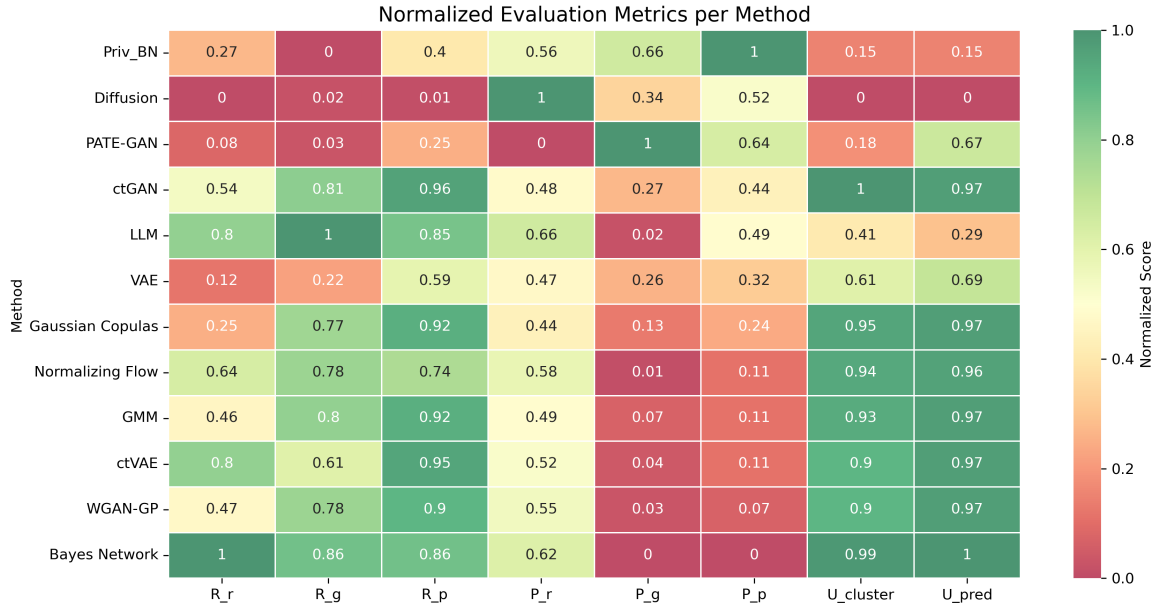

Figure 4: Normalized evaluation metrics per method including representativeness ($\mathcal{R}_r, \mathcal{R}_g, \mathcal{R}_p$), privacy ($\mathcal{P}_r, \mathcal{P}_g, \mathcal{P}_p$), and utility ($\mathcal{U}_{cluster}, \mathcal{R}_{pred}$).

Figure 5: Overall performance per dimension of benchmarking models.

Conditioned VAE (cVAE) demonstrates better overall performance than the unconditioned counterparts (VAE) as expected. However, the inclusion of additional side information in models like cVAE can lead to increased privacy leakage risks relative to VAE. Therefore, selecting an appropriate model requires careful consideration of the trade-off between performance and privacy in practical applications. Notably, the Normalizing Flow model achieved performance on par with cVAE, highlighting its strong potential as a foundation for further development in synthetic trip generation.

Interestingly, the LLM attains high representativeness but relatively low utility, diverging from the general trend observed in other models where utility scores are often aligned with representativeness. This suggests that preserving distributional similarity does not neces-

sarily guarantee strong downstream analytical performance. The results of LLM generator reinforce the importance of evaluating synthetic data beyond representativeness, as relying solely on distributional similarity metrics risks overestimating practical utility. These findings underscore again the value of a comprehensive, multi-dimensional evaluation framework.

It should be pointed out that to ensure comparability, all models in this study were trained on identically preprocessed data, in which categorical variables were one-hot encoded and continuous variables were min–max normalized. While this setting is consistent across models, it is not equally well-suited to every generative architecture. That is why diffusion model struggled. Diffusion models inject Gaussian noise directly into the feature space and iteratively denoise, which can disrupt the sparsity and mutual exclusivity inherent in one-hot representations, making it challenging to recover valid categorical structures.

5.3.2. Detailed comparative analysis

To provide a more concrete illustration of the evaluation framework, we conduct a detailed comparative analysis of four representative models: Bayes Network (BN), Priv-BN, CTGAN, and PATE-GAN. These models are selected as they span the key categories considered in this study: a statistical model (BN), a deep generative model (CTGAN), and their respective privacy-enhanced variants (Priv-BN and PATE-GAN). By comparing these models in greater detail, we highlight not only their relative performance across representativeness, privacy, and utility, but also the trade-offs that arise between fidelity and privacy when privacy-preserving mechanisms are incorporated.

Figure 6 shows the average performance per dimension for the four selected models, clearly illustrating the fundamental trade-off between utility and privacy: gains in one dimension often come at the expense of the other. Models with high utility, such as BN and CTGAN, tend to exhibit weaker privacy performance, indicating a greater risk of information leakage when the synthetic data closely mirrors the real data distribution. In contrast, privacy-enhanced models like Priv-BN and PATE-GAN achieve stronger privacy protection but generally at the cost of representativeness and downstream analytical utility.

Figure 7 presents the distribution of membership inference attack (MIA) scores for real training, holdout, and synthetic records across the four models. The kernel density estimation (KDE) curves illustrate how closely synthetic records resemble training or holdout data, which serves as an indicator of potential privacy leakage. The KDE distributions of synthetic records (green) overlap more strongly with those of the real training records (blue) than with holdout records (orange) for model BN and CTGAN. This suggests a higher risk of memorization and privacy leakage, as synthetic samples may reveal information about the training set. Compared to BN, the synthetic distribution of Priv-BN shifts slightly away from the training data, reducing overlap. This indicates that the differential privacy mechanism provides some mitigation of leakage, although at the expense of sample utility. The synthetic KDE of PATE-GAN is more distinct and separated from the training distribution, showing lower overlap with training data. This demonstrates stronger protection against membership inference, consistent with the model's design to incorporate privacy-preserving aggregation.

Figure 8 shows the kernel density estimation (KDE) of trip start times in the real dataset compared with synthetic data generated by the four models. BN and CTGAN capture the main distributional patterns of real trip start times well, with KDE curves closely aligned to the real data. This indicates strong representativeness at the distributional level. Priv-

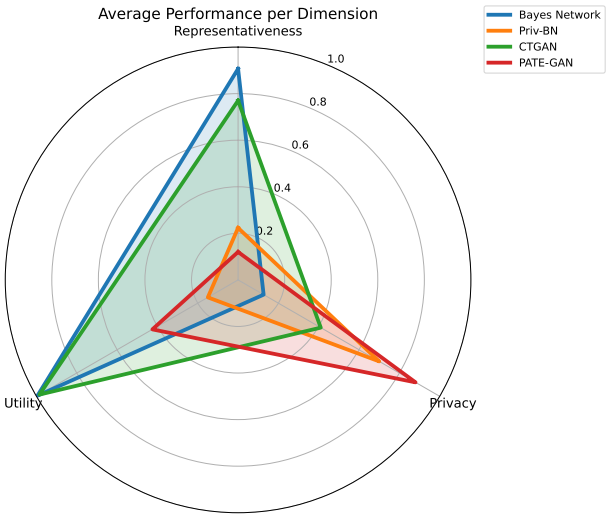

Figure 6: Average performance per dimension of Bayes Network, Priv-BN, CTGAN, and PATE-GAN.

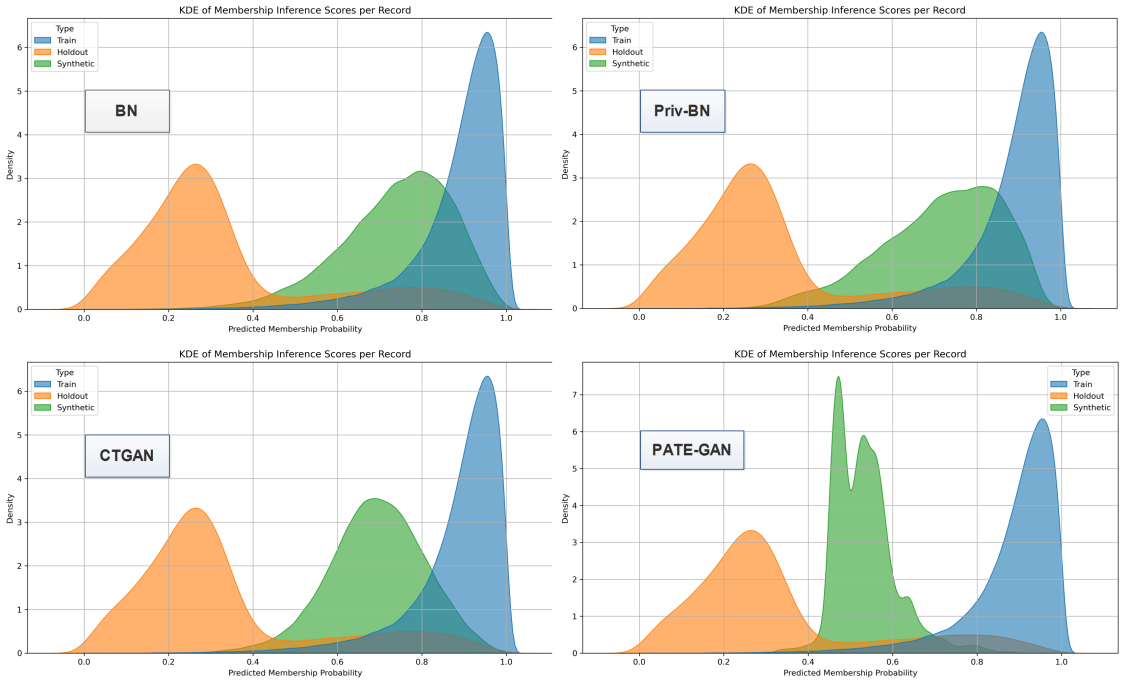

Figure 7: Privacy comparison of Bayes Network, Priv-BN, CTGAN, and PATE-GAN. The kernel density estimation (KDE) of membership inference attack (MIA) scores for synthetic records generated by each model are compared against the scores of real training and holdout (testing) records.

BN shows a clear deviation from the real distribution, with synthetic trips concentrated at unrealistic early start times, reflecting the utility loss introduced by privacy constraints. PATE-GAN similarly fails to capture the multi-modal structure of real start times, generating distributions concentrated in unrealistic ranges.

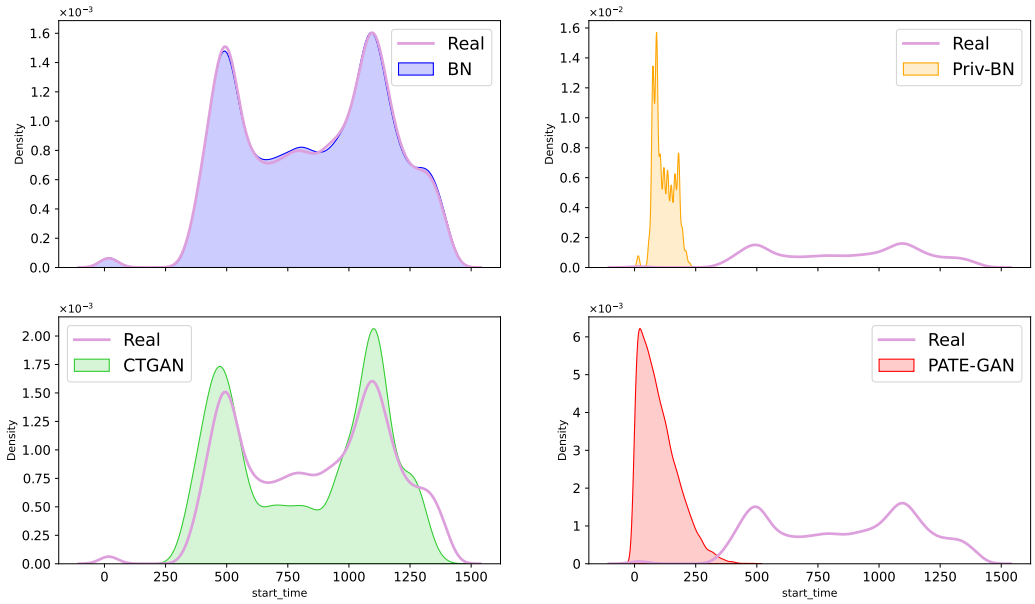

Figure 8: Representativeness comparison of Bayes Network, Priv-BN, CTGAN, and PATE-GAN. The kernel density estimates (KDE) of trip start times from the synthetic data generated by these models are compared against those from the real test data.

Figure 9 visualizes the cluster centroids from the real and synthetic datasets generated by the four models, projected into two dimensions using Principal Component Analysis (PCA). Consistent with the representativeness results, BN and CTGAN produce centroids that are closer to those of the real data, suggesting that their synthetic datasets are more suitable for clustering analysis. In contrast, the centroids from Priv-BN and PATE-GAN deviate substantially, indicating limited utility for clustering tasks.

The detailed comparison of BN, Priv-BN, CTGAN, and PATE-GAN reflects the broader patterns observed across all benchmarked methods. Taken together, the results underscore the importance of a multi-dimensional, multi-level evaluation when assessing generative models. While developed in the context of synthetic trips, the proposed framework is equally relevant to other applications such as trajectory generation, where outputs must also be assessed across multiple dimensions and levels. To support transparency and reproducibility, we make the complete preprocessing and training code available for reference and examination by other researchers.

6. Conclusion

This study introduced a comprehensive, multi-dimensional framework for evaluating synthetic trip data, spanning representativeness and privacy across record-, group-, and population-levels, as well as utility of different downstream tasks. Applying this framework to a diverse set of statistical and deep generative models revealed that no single approach dominates across all dimensions. Statistical models, such as Bayesian Networks, deliver strong repre-

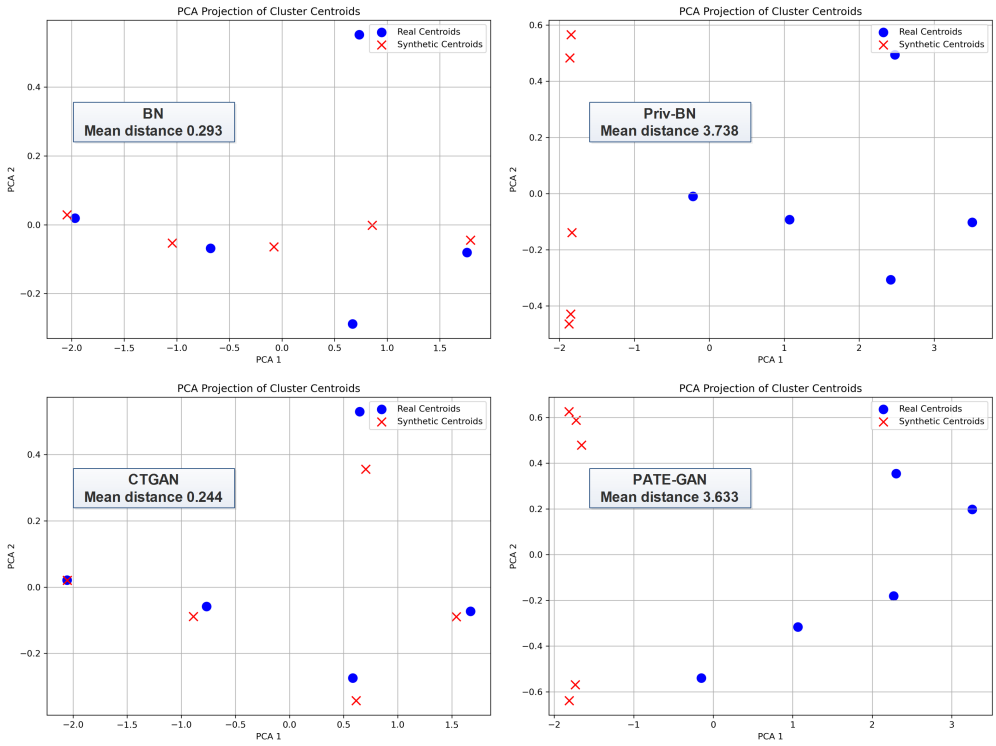

Figure 9: Utility comparison in the clustering task of Bayes Network, Priv-BN, CTGAN, and PATE-GAN. The evaluation is based on the average minimum distance between real and synthetic cluster centroids. Cluster centroids from the real and synthetic datasets are visualized in two dimensions using Principal Component Analysis (PCA).

sentativeness and utility but carry substantial privacy risks, while privacy-enhanced variants like Priv-BN offer improved protection at the cost of data fidelity and analytical value. Among deep generative models, CTGAN achieved the most balanced performance, demonstrating that integrating conventional modeling components, such as GMM-based sampling, can enhance both fidelity and privacy. In contrast, formal differential privacy mechanisms, exemplified by PATE-GAN, provided strong privacy guarantees but significantly reduced representativeness and utility.

Overall, our results highlight the inherent trade-offs between privacy and utility in synthetic data generation for public transit applications, and demonstrate the necessity of multi-dimensional, multi-level evaluation to avoid misleading conclusions from single-metric or single-level assessments. The proposed framework provides a transparent and reproducible basis for benchmarking generative models, guiding the selection and adaptation of methods for responsible deployment in transport research and policy analysis.

While the evaluation framework is broadly applicable, several limitations should be acknowledged. First, the results reflect a single real-world dataset with fixed preprocessing, which may influence model performance rankings. Second, the scope of evaluation, although multi-dimensional, does not exhaust all possible measures of privacy risk or domain-specific utility. Finally, hyperparameter tuning was conducted within practical constraints, and alternative settings might yield different trade-offs. Building on our current benchmarking findings, we plan to investigate hybrid architectures that integrate the strengths of statisti-

cal and deep generative approaches. We also will extend the framework to other transport datasets that contain more sensitive personal information.

Data and Code Availability

The code supporting the findings of this study is available at GitHub. Due to the sensitive nature of the data used in this study, the data will remain confidential and cannot be publicly shared. However, access may be granted upon reasonable request by contacting authors.

Acknowledgements

We would like to acknowledge TRENoP research centers in Sweden for their funding supports. This work was also in part financially supported by Digital Futures. During the preparation of this work the authors used ChatGPT 4o in order to check the English grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L., 2016. Deep learning with differential privacy, in: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp. 308–318.

Albrecht, T., Keller, R., Rebholz, D., Röglinger, M., 2024. Fake it till you make it: Synthetic data for emerging carsharing programs. Transportation research part D: transport and environment 127, 104067.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: International conference on machine learning, PMLR. pp. 214–223.

Badu-Marfo, G., Farooq, B., Patterson, Z., 2020. A differentially private multi-output deep generative networks approach for activity diary synthesis. arXiv preprint arXiv:2012.14574 .

Badu-Marfo, G., Farooq, B., Patterson, Z., 2022. Composite travel generative adversarial networks for tabular and sequential population synthesis. IEEE Transactions on Intelligent Transportation Systems 23, 17976–17985.

Bao, J., Li, L., Davis, A., 2022. Variational autoencoder or generative adversarial networks? A comparison of two deep learning methods for flow and transport data assimilation. Mathematical Geosciences 54, 1017–1042.

Berke, A., Doorley, R., Larson, K., Moro, E., 2022. Generating synthetic mobility data for a realistic population with rnns to improve utility and privacy, in: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, pp. 964–967.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., Kasneci, G., 2022. Language models are realistic tabular data generators. arXiv preprint arXiv:2210.06280 .

Borysov, S.S., Rich, J., Pereira, F.C., 2019. How to generate micro-agents? a deep generative modeling approach to population synthesis. Transportation Research Part C: Emerging Technologies 106, 73–97.

Choi, S., Jin, Z., Ham, S.W., Kim, J., Sun, L., 2025. A gentle introduction and tutorial on deep generative models in transportation research. Transportation Research Part C: Emerging Technologies 176, 105145.

Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M., 2023. Diffusion models in vision: A survey. IEEE transactions on pattern analysis and machine intelligence 45, 10850–10869.

Dinh, L., Sohl-Dickstein, J., Bengio, S., 2016. Density estimation using real nvp. arXiv preprint arXiv:1605.08803 .

Doersch, C., 2016. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 .

Drechsler, J., 2011. Synthetic datasets for statistical disclosure control: theory and implementation. volume 201. Springer Science & Business Media.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Calibrating noise to sensitivity in private data analysis, in: Theory of cryptography conference, Springer. pp. 265–284.

Eigenschink, P., Reutterer, T., Vamosi, S., Vamosi, R., Sun, C., Kalcher, K., 2023. Deep generative models for synthetic data: A survey. IEEE Access 11, 47304–47320.

Fontana, I., Langheinrich, M., Gjoreski, M., 2023. Gans for privacy-aware mobility modeling. IEEE Access 11, 29250–29262.

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., Sales, A.P., 2020. Generation and evaluation of synthetic patient data. BMC medical research methodology 20, 108.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems 27.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein gans. Advances in neural information processing systems 30.

Halvorsen, A., Koutsopoulos, H.N., Ma, Z., Zhao, J., 2020. Demand management of congested public transport systems: a conceptual framework and application using smart card data. Transportation 47, 2337–2365.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840–6851.

Huang, D., Song, X., Fan, Z., Jiang, R., Shibasaki, R., Zhang, Y., Wang, H., Kato, Y., 2019. A variational autoencoder based generative model of urban human mobility, in: 2019 IEEE conference on multimedia information processing and retrieval (MIPR), IEEE. pp. 425–430.

Jordon, J., Yoon, J., Van Der Schaar, M., 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees, in: International conference on learning representations.

Jutras-Dubé, P., Al-Khasawneh, M.B., Yang, Z., Bas, J., Bastin, F., Cirillo, C., 2024. Copula-based transferable models for synthetic population generation. Transportation Research Part C: Emerging Technologies 169, 104830.

Kapp, A., Hansmeyer, J., Mihaljević, H., 2023. Generative models for synthetic urban mobility data: A systematic literature review. ACM Computing Surveys 56, 1–37.

Kieu, M., Meredith, I.B., Raith, A., 2023. Synthetic generation of trip data: The case of smart card. Data Science for Transportation 5, 14.

Kim, E.J., Bansal, P., 2023. A deep generative model for feasible and diverse population synthesis. Transportation Research Part C: Emerging Technologies 148, 104053.

Kim, E.J., Kim, D.K., Sohn, K., 2022. Imputing qualitative attributes for trip chains extracted from smart card data using a conditional generative adversarial network. Transportation Research Part C: Emerging Technologies 137, 103616.

Kingma, D.P., Welling, M., et al., 2013. Auto-encoding variational bayes.

Koller, D., Friedman, N., 2009. Probabilistic graphical models: principles and techniques. MIT press.

Kulkarni, V., Tagasovska, N., Vatter, T., Garbinato, B., 2018. Generative models for simulating mobility trajectories. arXiv preprint arXiv:1811.12801 .

Lautrup, A.D., Hyrup, T., Zimek, A., Schneider-Kamp, P., 2025. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. Data Mining and Knowledge Discovery 39, 6.

Lee, H., Bansal, P., Vo, K.D., Kim, E.J., 2025. Collaborative generative adversarial networks for fusing household travel survey and smart card data to generate heterogeneous activity schedules in urban digital twins. Transportation Research Part C: Emerging Technologies 176, 105125.

Li, L., Bi, J., Yang, K., Luo, F., 2022. Spatial-temporal semantic generative adversarial networks for flexible multi-step urban flow prediction, in: International Conference on Artificial Neural Networks, Springer. pp. 763–775.

Lim, S.Y., Yun, H., Bansal, P., Kim, D.K., Kim, E.J., 2025. A large language model for feasible and diverse population synthesis. arXiv preprint arXiv:2505.04196 .

Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O., 2018. Are gans created equal? a large-scale study. Advances in neural information processing systems 31.

Ma, Z., Koutsopoulos, H.N., Liu, T., Basu, A.A., 2020. Behavioral response to promotion-based public transport demand management: Longitudinal analysis and implications for optimal promotion design. Transportation Research Part A: Policy and Practice 141, 356–372.

Mo, B., Ma, Z., Koutsopoulos, H.N., Zhao, J., 2021. Calibrating path choices and train capacities for urban rail transit simulation models using smart card and train movement data. Journal of Advanced Transportation 2021, 5597130.

Nassir, N., Hickman, M., Ma, Z.L., 2019. A strategy-based recursive path choice model for public transit smart card data. Transportation Research Part B: Methodological 126, 528–548.

Nelsen, R.B., 2006. An introduction to copulas. Springer.

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B., 2021. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research 22, 1–64.

Patel, A., Raffel, C., Callison-Burch, C., 2024. Datadreamer: A tool for synthetic data generation and reproducible llm workflows. arXiv preprint arXiv:2402.10379 .

Patki, N., Wedge, R., Veeramachaneni, K., 2016. The synthetic data vault, in: 2016 IEEE international conference on data science and advanced analytics (DSAA), IEEE. pp. 399–410.

Pelletier, M.P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, 557–568.

Peng, M., Chen, K., Guo, X., Zhang, Q., Lu, H., Zhong, H., Chen, D., Zhu, M., Yang, H., 2024. Diffusion models for intelligent transportation systems: A survey. arXiv preprint arXiv:2409.15816 .

Platzer, M., Reutterer, T., 2021. Holdout-based fidelity and privacy assessment of mixed-type synthetic data. arXiv preprint arXiv:2104.00635 .

Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S., 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE access 12, 26839–26874.

Reynolds, D., 2015. Gaussian mixture models, in: Encyclopedia of biometrics. Springer, pp. 827–832.

Rong, C., Ding, J., Liu, Z., Li, Y., 2023a. Complexity-aware large scale origin-destination network generation via diffusion model. arXiv preprint arXiv:2306.04873 .

Rong, C., Wang, H., Li, Y., 2023b. Origin-destination network generation via gravity-guided gan. arXiv preprint arXiv:2306.03390 .

Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), IEEE. pp. 3–18.

Smolak, K., Rohm, W., Knop, K., Siła-Nowicka, K., 2020. Population mobility modelling for mobility data simulation. Computers, Environment and Urban Systems 84, 101526.

Sweeney, L., 2002. k-anonymity: A model for protecting privacy. International journal of uncertainty, fuzziness and knowledge-based systems 10, 557–570.

Uğurel, E., Huang, S., Chen, C., 2024. Learning to generate synthetic human mobility data: A physics-regularized gaussian process approach based on multiple kernel learning. Transportation Research Part B: Methodological 189, 103064.

Vo, K.D., Kim, E.J., Bansal, P., 2025. A novel data fusion method to leverage passively-collected mobility data in generating spatially-heterogeneous synthetic population. Transportation Research Part B: Methodological 191, 103128.

Wang, S., Cao, J., Chen, H., Peng, H., Huang, Z., 2020. SeqST-GAN: Seq2seq generative adversarial nets for multi-step urban crowd flow prediction. ACM Transactions on Spatial Algorithms and Systems (TSAS) 6, 1–24.

Wu, Y., Markham, A., Wang, L., Solus, L., Ma, Z., 2025. Data-driven causal behaviour modelling from trajectory data: A case for fare incentives in public transport. Journal of Public Transportation 27, 100114.

Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K., 2019. Modeling tabular data using conditional gan. Advances in neural information processing systems 32.

Xu, X., Ye, Z., Zhang, A., Liu, J., Wu, Y., Qin, L., Xu, W., Ran, B., 2024. A generative adversarial network for inferring age attributes in subway based on automatic fare collection and point of interest data. Journal of Intelligent Transportation Systems , 1–21.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H., 2023. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys 56, 1–39.

Zhang, J., Cormode, G., Procopiuc, C.M., Srivastava, D., Xiao, X., 2017. Privbayes: Private data release via bayesian networks. ACM Transactions on Database Systems (TODS) 42, 1–41.

Zhang, M., Lin, H., Takagi, S., Cao, Y., Shahabi, C., Xiong, L., 2023. CSGAN: Modality-aware trajectory generation via clustering-based sequence GAN, in: 2023 24th IEEE International Conference on Mobile Data Management (MDM), IEEE. pp. 148–157.

Zhang, Y., Yuan, Y., Ding, J., Yuan, J., Li, Y., 2025. Noise matters: Diffusion model-based urban mobility generation with collaborative noise priors, in: Proceedings of the ACM on Web Conference 2025, pp. 5352–5363.

Zhao, Z., Birke, R., Chen, L.Y., 2025. Tabula: Harnessing language models for tabular data synthesis, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer. pp. 247–259.