Privacy-Preserving Data Analytics Using Synthetic Data Generation

Final Report

by

Murat Kantarcioglu
University of Texas at Dallas

Alvaro Cardenas, University of California, Santa Cruz
Latifur Khan, University of Texas at Dallas
Bhavani Thuraisingham, University of Texas at Dallas
Gurcan Comert, Benedict College

May 2025

NATIONAL CENTER FOR TRANSPORTATION CYBERSECURITY AND RESILIENCY (TraCR)

DISCLAIMER

CONTACTS

For more information:

Murat Kantarcioglu
Data and Decision Sciences, RM 386
727 Prices Fork Rd.
Blacksburg, VA 24061
Phone: 972-883-6616/(540) 231-6772
Email:muratk@utdallas.edu,
muratk@vt.edu

TraCR
Clemson University
One Research Dr
Greenville, SC 29607
tracr@clemson.edu

ACKNOWLEDGMENT

This work is based upon the work supported by the National Center for Transportation Cybersecurity and Resiliency (TraCR), a U.S. Department of Transportation National University Transportation Center headquartered at Clemson University, Clemson, South Carolina, USA. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of TraCR. The U.S. Government assumes no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No. 6	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Privacy-Preserving Data Analytics Using Synthetic Data Generation.		**5. Report Date:** May 2025 **6. Performing Organization Code:** N/A
7. Author(s) Murat Kantarcioglu, Ph.D.; https://orcid.org/0000-0001-9795-9063 Alvaro Cardenas, Ph.D.; https://orcid.org/0000-0002-5142-9750 Latifur Khan, Ph.D.; https://orcid.org/0000-0002-9300-1576 Bhavani Thuraisingham, Ph.D.; https://orcid.org/0000-0003-4653-2080 Gurcan Comert, Ph.D.; https://orcid.org/0000-0002-2373-5013		**8. Performing Organization Report No.** 6
9. Performing Organization Name and Address National Center for Transportation Cybersecurity and Resiliency (TraCR), Clemson University, 414 A One Research Dr, Greenville, SC 29607 University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080		**10. Work Unit No.** N/A **11. Contract or Grant No.** 69A3552344812 and 69A3552348317
12. Sponsoring Agency Name and Address U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology, 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Report, 01/01/2024 - 12/31/2024 **14. Sponsoring Agency Code** OST-R
15. Supplementary Notes Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.		

16. Abstract

The sharing of large-scale transportation data is beneficial for transportation planning and policymaking. However, it also raises significant security and privacy concerns, as the data may include identifiable personal information, such as individuals' home locations. To address these concerns, synthetic data generation based on real transportation data offers a promising solution that allows privacy protection while potentially preserving data utility. Although there are various synthetic data generation techniques, they are often not tailored to the unique characteristics of transportation data, such as the inherent structure of transportation networks formed by all trips in the datasets. In this paper, we use New York City taxi data as a case study to conduct a systematic evaluation of the performance of widely used tabular data generative models. In addition to traditional metrics such as distribution similarity, coverage, and privacy preservation, we present a novel graph-based metric tailored specifically for transportation data. This metric evaluates the similarity between real and synthetic transportation networks, providing potentially deeper insights into their structural and functional alignment. We also introduced an improved privacy metric to address the limitations of the commonly used one. Our experimental results reveal that existing tabular data generative models often fail to perform as consistently as claimed in the literature, particularly when applied to transportation data use cases. Furthermore, our novel graph metric reveals a significant gap between synthetic and real data. This work underscores the potential need to develop generative models specifically tailored to take advantage of the unique characteristics of different domains, such as transportation.

17. Keywords Tabular Data Synthesis · Transportation Data · Generative Models Evaluation	**18. Distribution Statement** No restrictions.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 23	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

List of Tables

List of Figures

EXECUTIVE SUMMARY

The sharing of large-scale transportation data is beneficial for improving transparency, traffic management, urban planning and policymaking. However, it also raises significant security and privacy concerns, as human mobility data is highly sensitive and may include identifiable personal information, such as individuals' home locations, medical visits, and place of worship, which may be abused by malicious parties.

To address these concerns, synthetic data generation based on real transportation data offers a promising solution that provides potentially high utility for downstream tasks while effectively mitigating privacy concerns. Although there are various synthetic data generation techniques, especially the deep learning-based models, such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models, they are often not tailored to the unique characteristics of transportation data, such as the inherent structure of transportation networks formed by all trips in the datasets.

In this project, we use New York City taxi data as a case study to conduct a systematic evaluation of the performance of widely used tabular data generative models. The New York City taxi data is one of the largest publicly available datasets in the domain. Its extensive size and accessibility make it a valuable resource for studies in this field. It is represented in tabular data format, with each row containing detailed information about an individual taxi trip.

We use a comprehensive set of metrics to evaluate the performance of various tabular data generative models, including Gaussian Copula, CTGAN, TVAE, CTABGAN, and two diffusion-based models: TabDDPM and STaSy. Our evaluation framework includes metrics such as downstream task performance (e.g., training predictive models on synthetic data to estimate taxi fares), Wasserstein distance, coverage, privacy preservation, and model complexity. Additionally, we introduce a novel graph-based metric and an enhanced privacy preservation metric to provide a more thorough assessment.

The novel graph-based metric is a metric tailored particularly for transportation data, as transportation data form a transportation graph, with origins and destinations as nodes, and each trip as an edge between two nodes. This metric evaluates the similarity between real and synthetic transportation networks, measured by the total variation distance, as each graph corresponds to a probability distribution. This metric provides potentially deeper insights into their structural and functional alignments.

The improved privacy measurement metric is presented based on the insight that the distance between synthetic data and training data is likely to be smaller than the distance between synthetic data and testing data if there is overfitting in the trained generative models. This addresses the shortcoming of the commonly used privacy measurement metric, Distance to Closest Record, which does not differentiate between these two sets of distances.

Our experimental results reveal that the existing tabular data generative models often fail to perform as consistently as claimed in the literature, specifically when applied to transportation

data. The results indicate that TabDDPM achieves the best overall performance across multiple metrics. However, it appears TabDDPM may struggle to handle categorical variables with hundreds of classes. Furthermore, our novel graph-based utility measurement metric reveals a significant gap between synthetic and real data. This work underscores the potential need to develop generative models specifically tailored to take advantage of the unique characteristics of transportation applications.

CHAPTER 1

Introduction

Open large-scale transportation data offer significant benefits, including enhancing transparency, improving traffic management and urban planning, and providing valuable opportunities for researchers to conduct in-depth studies that inform and support effective policymaking. However, the availability of largescale data raises significant privacy concerns, as human mobility is highly sensitive. Transportation data may contain identifiable personal information such as individuals' home locations, and data sharing can infringe on individual privacy. For example, celebrities may be stalked with shared taxi journey data [2].

To address these challenges, synthetic data generation techniques have emerged as a promising solution for data sharing, offering potentially high utility for downstream tasks while effectively mitigating the privacy concerns. These techniques essentially approximate the raw data distribution with machine learning models and then generate artificial data instead of sharing the raw data directly.

In recent years, deep learning-based synthetic data generation models have drawn considerable attention due to their ability to learn complex data distributions and generate realistic synthetic data. Among deep generative models, Generative Adversarial Networks (GANs) [8], Variational Autoencoders (VAEs) [13], and diffusion models [10] have demonstrated remarkable capabilities in generating high-quality samples, particularly in domains such as images and text.

Deep generative models have also been adapted for various data applications, including tabular data with structured rows and columns [27]. This prevalent form of data representation is widely used in diverse domains, including finance, healthcare, and e-Commerce, among others [4]. Tabular data is also a common format for sharing transportation data. For instance, taxi trips can be represented in a tabular format, where each row corresponds to an individual trip, and the columns capture various properties of the trip, such as the start location, end location, and trip duration.

Researchers have proposed a variety of models to generate synthetic tabular data, such as GAN and Variational autoencoder-based, diffusion-based generative modeling [14,27]. To synthesize transportation data, leveraging existing tabular data generative models provides an efficient starting point for further research. However, it is observed that current generative models are primarily designed for general tabular data synthesis, leaving several important questions unanswered:

1. Transportation data possesses unique properties; for instance, the data collectively forms a transportation network. Can the existing metrics effectively capture and measure this characteristic?
2. Tabular data generative models typically do not prioritize privacy preservation. How effectively do they perform when evaluated from a privacy preservation perspective?

3. The datasets commonly used to evaluate these models generally do not include transportation data. Will the performance ranking of these models remain consistent when applied to transportation data?

In this paper, we aim to answer these questions by conducting a systematic evaluation of typical tabular data generative models [12,14,19,27,28] on transportation data, using New York City taxi data [25] as a case study. Our main contributions can be summarized as follows:

1. We propose a novel graph-based metric to quantify the property gap between real and synthetic transportation data, leveraging the fact that transportation networks can naturally be represented as graphs.
2. We propose an improved privacy leakage metric to investigate the privacy-preserving capabilities of these models and assess their vulnerabilities, particularly to membership inference attacks [21].
3. We perform a systematic evaluation of the performance of these models within the transportation domain, utilizing a range of metrics such as downstream task utility, distribution similarity, and diversity.

In the remainder of this paper, we first provide a brief review of related work in Section 2 and present the necessary background knowledge in Section 3 to facilitate a better understanding of our work. Next, we introduce novel evaluation metrics in Section 4. The experimental setup and results are detailed in Section 5, followed by a concluding summary in Section 6.

CHAPTER 2
Related Work

2.1 Generative Models

Generative Adversarial Networks (GANs) [8] are among the most widely used generative models. It employs a generator and a discriminator, two competing neural networks; the generator tries to trick the discriminator to classify the fake data as real, while the discriminator tries to differentiate real and fake data.

Variational Autoencoders (VAEs) [13] are another class of generative models, which map real data to a distribution within a latent space by an encoder, then a decoder maps from the latent space to the input space.

Diffusion models represent the latest advancement in generative models [10]. They involve a forward diffusion process and a reverse denoising process. In the forward process, noise is gradually added to the training data with increasing magnitude until the data becomes pure noise. In the reverse process, a model is trained to denoise the noisy data, effectively reconstructing the clean data and learning the underlying data distribution.

2.2 Tabular Data Generation

Tabular data pose unique challenges for synthetic data generation. Unlike image data, tabular data often consist of a mix of continuous and discrete variables. Moreover, values in the discrete columns frequently exhibit imbalanced distributions, adding an additional layer of complexity to the generation process. [27] proposes a conditional tabular GAN (CTGAN) to address these challenges. CTGAN employs two distinct sampling approaches to handle discrete and continuous variables in the training data. For discrete variables, it first randomly selects a discrete column, then samples rows based on the logarithm frequency of categorical values in that column. The sampled categorical values will serve as conditional inputs to GAN. For continuous variables, it estimates the number of modes for each column with variational Gaussian mixture models [3] and samples by modes and normalizes the values. [27] also proposed tabular VAE (TVAE) by adapting VAE to tabular data.

[28] makes improvements upon CTGAN motivated by several observations: *Within one variable* of the tabular data there may be mixed continuous and categorical data types, and its distribution may be skewed and have a long tail. The authors address these issues by proposing mode-value pair for mixed data types, logarithmic transformation for variables with long tail distribution, and an additional continuous mode as the conditional input to GAN.

[14] proposed TabDDPM by adapting diffusion models to the tabular data domain, employing Gaussian diffusion models for continuous variables and multinomial diffusion models for categorical variables [11]. [12] proposed the STaSy model by directly adapting score-based generative modeling [22] to the tabular data domain.

While these techniques have shown promise in tabular data, to the best of our knowledge, they have not been evaluated in the context of transportation data with their unique characteristics. In this work, we evaluate the aforementioned tabular synthetic data generation techniques within the context of a transportation data use case.

CHAPTER 3

Background

3.1 Evaluation Metrics for Generative Models

The quality of synthetic data is evaluated based on its utility and its ability to preserve privacy. However, higher privacy protection can potentially reduce a model's utility. Conversely, higher utility (i.e., synthetic data that closely resembles real data) may increase the privacy risks. Balancing utility and privacy in synthetic data generation is challenging and remains an active research area
[23].

In this context, utility can be assessed through various measures, including downstream task performance, statistical similarity, and diversity. Privacy preservation can be evaluated by measuring the distance between a synthetic data point and real data. Below, we present a brief description of these measures.

Downstream Task Performance To evaluate the utility of the generated synthetic data, we employ a selected downstream task. In the context of taxi ride information, one critical piece of information is the total cost of the ride. The key question, therefore, is how effectively the synthetic data can be used to train a machine learning model capable of accurately predicting the total cost of a taxi ride. More specifically, we generate synthetic data with the trained generative model, train a prediction model "Gradient Boosting for Regression" [16] with the synthetic data, then we predict the "total amount" (i.e., the total amount paid for the taxi ride) with the training data and synthetic data respectively. The performance is represented by coefficient of determination [15]. The best possible value of R^2 is 1.

Similarity We use Wasserstein distance [1] to measure the similarity between two distributions (i.e., real vs synthetic data).

Diversity We use coverage [18] to measure the diversity of a distribution, enabling us to assess whether mode collapse [8] has occurred. Coverage is calculated as the percentage of real sample hyperspheres which contain a generated sample. The real sample hypersphere is calculated with its K^{th} nearest neighbor. It is found to be more robust than the metric recall [26].

Privacy Measure The distance of a synthetic data point to its closest real data neighbor (DCR) serves as a metric for evaluating privacy preservation in synthetic data generation [28]. This ensures that synthetic records are not overly similar to individual records in the original dataset, thereby reducing the risk of privacy breaches. This metric is also closely related to membership inference attacks [21], where a distance-based metric [9] is utilized to determine whether a data point was included in the training dataset of the model under attack. We explore this connection in greater detail in Section 4.2.

CHAPTER 4

The Novel Evaluation Metrics

In this section, to evaluate the generated synthetic data for transportation applications, we propose two novel metrics. To the best of our knowledge, these metrics have not been previously used in the context of evaluating synthetic tabular transportation data generation.

4.1 Graph Similarity Metric for Transportation Network Data

Transportation data, when viewed collectively, forms a transportation network. This network can be effectively represented as a graph, capturing the overall transportation trends and relationships within the data. For instance, the pickup and drop-off locations in the NYC taxi dataset correspond to different zones within the city. These zones can be represented as nodes in a graph, providing a structured way to model the transportation network. In other words, each trip between two zones can be represented as an edge in the transportation graph. Let the number of trips between two zones i and j be n_{ij}, then the total number of trips $N = \sum_{i,j} n_{ij}$. Let the fraction of edges between two nodes i and j of the transportation graph G be $p_G(i,j) = \frac{n_{ij}}{N}$. Clearly, $\sum_{i,j} p_G(i,j) = 1$, which means the fraction of edges p_G represents a distribution.

We can construct a transportation graph from the real transportation data, denoted as G_r, and another graph, G_s, from the generated synthetic transportation data. We can measure the similarity between the real transportation graphs G_r and the synthetic ones G_s by calculating the similarity score S_G between the two graphs as follows:

$$S_G(G_r, G_s) = 1 - \delta(p_{G_r}, p_{G_s})$$
$$s.t. \quad \delta(p_{G_r}, p_{G_s}) = \frac{1}{2} \sum_{i,j} |p_{G_r}(i,j) - p_{G_s}(i,j)|, \tag{1}$$

where p_{G_r} and p_{G_s} represent edge number distributions for the real and synthetic graphs G_r, G_s respectively, and $\delta(p_{G_r}, p_{G_s})$ is the total variation distance.

4.2 Distance to Closest Record Ratio as Privacy Leakage Metric

The success of membership inference attacks relies on the observation that models tend to overfit their training data. Consequently, the distance between *training* data and synthetic data is smaller than the distance between *testing* data and synthetic data. This phenomenon is also evidenced by the fact that training data loss is typically smaller than testing data loss. Based on this observation, relying solely on the distance between real data and synthetic data, as widely used in previous literature [5,28], may be insufficient to reliably assess the risk of privacy leakage.

We therefore propose a more robust metric that uses two distances instead of one, comparing the two distances by calculating their ratio. Specifically, we set aside holdout *testing* data besides the real *training* data. Let the distance of real data to the closest synthetic data be $d_\alpha(r,s)$, and the distance of holdout testing data to the closest synthetic data be $d_\alpha(h,s)$, where α is a percentile of

all the closest distance values. The Distance to Closest Record Ratio (rDCR) is defined as $r_{DCR} = \frac{d_\alpha(r,s)}{d_\alpha(h,s)}$.

When $r_{DCR} < 1$, the distance between the training data and synthetic data is smaller than the distance between testing and synthetic data. This indicates overfitting, making the model vulnerable to membership inference attacks. A smaller ratio r_{DCR} indicates greater overfitting of the model to the training data, making the model more vulnerable to a potential membership inference attack. On the other hand, if $r_{DCR} > 1$, a small distance of $d_\alpha(r,s)$ alone may not be sufficient to demonstrate the vulnerability of a model to distanced-based membership inference attacks.

The metric from [20] bears a resemblance to the rDCR metric described in this work, but there are significant differences. Our metric focuses on the distance to the closest synthetic data for each real data as it is designed to analyze the privacy leakage for the real data. In contrast, their metric measures the closest distance to real data for each synthetic data. Additionally, we incorporate a percentile-based approach to assess privacy leakage, recognizing that typically only a small subset of training data is vulnerable to membership inference attacks [6]. By examining the percentile of data at risk of privacy leakage, this percentile-based approach provides a more refined method for evaluating privacy leakage.

CHAPTER 5

Experiments

5.1 Datasets

In this work, we use New York City taxi trip data [24] as the experimental dataset. This dataset is widely utilized in transportation research [7,17] and is one of the largest publicly available datasets in the domain. Its extensive size and accessibility make it a valuable resource for studies in this field. It is represented in a tabular data format, with each row containing detailed information about an individual taxi trip.

More specifically, we use "2015 Green Taxi Trip Data" [25]. It has 19.2 million rows and each row has 21 columns. Each row represents a single trip in a green taxi, and the column fields include location and time for both pickup and dropoff, trip distance, itemized fares, payment type, tax and passenger count etc.

We pre-process the data by dropping columns 'Ehail fee' due to too many 'NaN' values, changing each pickup/drop-off column 'datetime' into two columns 'weekday' and 'time'. The final data has 22 columns with 8 categorical variables, 2 integer variables and 12 float-type numerical variables. This transportation dataset is more complicated than the tabular datasets usually used in the previous tabular data generative model papers, due to its larger data size, mixed data type and higher dimensionality. In the experiments, we randomly sample a subset of the data: the test dataset size is 20,000, and the training data size is 40,000 by default, unless specified otherwise. The proposed graph metric requires knowledge of the zones into which the pickup and drop-off longitude and latitude coordinates fall for each trip. To fulfill this requirement, we utilize the New York City Green Taxi Trip Records from March 2019. A key distinction of this dataset, compared to the aforementioned one, is its less granular nature: locations are represented by zones rather than precise longitude and latitude coordinates. Notably, the zone variable is categorical, unlike longitude and latitude, which are numerical. This difference makes the dataset particularly well-suited for zone-based transportation metrics.

5.2 Generative Models Used for Evaluation

The generative models evaluated in this paper include Gaussian Copula [19], CTGAN and TVAE [27], CTABGAN [28] and two diffusion-based tabular data generative models: TabDDPM [14] and STaSy [12]. We evaluate these models using various metrics including utility, similarity, diversity and privacy leakage as detailed in Sections 3.1 and 4.

5.3 Experimental Setup

During the experiments, for each method, three models are trained, and five times of sampling are conducted for each trained model. We limit the sample size to 20,000 for each sampling iteration due to the high memory demands of the following Wasserstein distance calculations.

The results are reported as the mean and standard deviation across a total of 15 sampling iterations.

5.4 Experimental Results

We report the results in separate tables, with bold fonts to highlight the best performance values, except for columns with values provided for reference purposes. All columns without synthetic data involved are for reference purposes.

Table 1: In the downstream task performance testing, the model predicts the "total amount". The R^2 values are multiplied by 100.

model	dwn tr tr	dwn tr syn	dwn tr te	dwn syn syn	dwn syn tr	dwn syn te
GaussianCopula	99.93 (0.00)	75.01 (2.90)	98.80 (0.33)	99.33 (0.01)	77.55 (1.20)	80.64 (0.23)
CTGAN	99.93 (0.01)	54.34 (1.19)	98.86 (0.42)	69.41 (1.40)	78.05 (2.29)	80.57 (2.21)
TVAE	99.93 (0.01)	**82.42 (2.00)**	98.84 (0.30)	90.03 (0.76)	72.32 (2.41)	74.27 (2.83)
CTABGAN	99.93 (0.00)	2.34 (23.57)	98.75 (0.35)	46.28 (3.09)	54.71 (20.50)	53.93 (22.83)
stasy	99.93 (0.00)	54.93 (6.97)	98.68 (0.41)	76.87 (3.78)	88.20 (4.45)	88.35 (4.74)
TabDDPM	99.93 (0.00)	60.26 (20.20)	98.92 (0.29)	89.38 (4.77)	**94.58 (1.38)**	**94.69 (1.55)**

Downstream Task Performance As described in Section 3.1, the model predicts the total amount for a given trip based on the other trip information. The results are reported in Table 1, where "model" is the model name, "dwn tr tr" means training on training data and predicting on training data, "dwn tr syn" means training on training data and predicting on synthetic data, "dwn tr te" means training on training data and predicting on testing data, provided for reference purpose, "dwn syn syn" means training on synthetic data and predicting on synthetic data, "dwn syn tr" means training on synthetic data and predicting on training data, and "dwn syn te" means training on synthetic data and predicting on testing data. The three columns "dwn tr syn", "dwn syn tr" and "dwn syn te" demonstrate the performance of these models. The performance of the models trained on synthetic data is particularly critical, as in real-world applications of data synthesis, only the synthetic data is typically made public for downstream tasks. The R^2 values in the table are multiplied by 100. Our results demonstrate that TabDDPM achieves the best downstream task performance among the evaluated methods.

Table 2: The Wasserstein distances.

model	w1 tr te	w1 tr syn	w1 te syn
GaussianCopula	0.1365 (0.0062)	1.0357 (0.0935)	1.0340 (0.0961)
CTGAN	0.1403 (0.0129)	0.7078 (0.0841)	0.6991 (0.0803)
TVAE	0.1262 (0.0046)	0.9093 (0.1090)	0.9057 (0.1142)
CTABGAN	0.1436 (0.0092)	**0.4260 (0.0193)**	**0.4306 (0.0216)**
stasy	0.1243 (0.0035)	1.0418 (0.0437)	1.0328 (0.0357)
TabDDPM	0.1230 (0.0025)	**0.4421 (0.0326)**	**0.4442 (0.0331)**

Statistical Similarity We report the experimental results for the Wasserstein distances in Table 2, where "w1 tr te" is the Wasserstein distance between the training data and the testing data, provided for reference purpose, "w1 tr syn" is the Wasserstein distance between the training data and the synthetic data, and "w1 te syn" is the Wasserstein distance between the testing data and the synthetic data. The results demonstrate that CTABGAN and TabDDPM have the best performance among all the models.

Graph Similarity Metric We report the graph similarity results in Table 3, where "G tr _te" is the graph similarity between the training data and the testing data, provided for reference purpose, "G tr syn" is the graph similarity between the training data and the synthetic data, "G te syn" is the graph similarity score between the testing data and the synthetic data. The similarity values are multiplied by 100 in the table. The results indicate that all models exhibit a significant performance gap compared to the reference value 73.17 given in column "G tr te". The TabDDPM model shows particularly low graph metric values. Further investigation shows that TabDDPM suffers severe mode collapse. We speculate that this issue may arise from its difficulty in handling categorical variables with hundreds of classes, such as the "zone" variable, or it may require significant additional hyperparameter tuning. Note that here we use the dataset with zone-based locations, as mentioned in Section 5.1. The results for the STaSy model are unavailable due to out-of-memory related issues, likely caused by challenges in achieving convergence.

Table 3: The graph similarity score, where the original values are multiplied by 100.

model	G tr te	G tr syn	G te syn
GaussianCopula	73.17 (0.00)	28.56 (0.87)	27.21 (0.31)
CTGAN	73.17 (0.00)	25.87 (0.72)	24.69 (0.38)
TVAE	73.17 (0.00)	**33.21 (2.41)**	**32.67 (2.04)**
CTABGAN	73.17 (0.00)	**32.12 (6.09)**	**29.96 (6.06)**
STaSy	73.17 (0.00)	N/A	N/A
TabDDPM	73.17 (0.00)	11.56 (2.84)	11.17 (2.80)

Table 4: The Coverage in the table is reported as the percentage of the coverage.

model	cov tr te	cov tr syn	cov te syn
GaussianCopula	74.60 (0.24)	0.62 (0.12)	0.65 (0.13)
CTGAN	74.88 (0.17)	19.95 (1.40)	19.52 (1.33)
TVAE	74.78 (0.17)	13.94 (0.48)	13.79 (0.43)
CTABGAN	74.83 (0.15)	2.96 (0.13)	3.02 (0.11)
stasy	74.80 (0.12)	2.34 (0.34)	2.40 (0.32)
TabDDPM	75.01 (0.44)	**68.69 (0.61)**	**67.92 (0.69)**

Diversity We report the percentage of the coverage in Table 4, where "cov tr te" is the coverage of the training data by the testing data, provided for reference purpose, "cov _tr syn" is the coverage of the training data by synthetic data and "cov _te syn" is the coverage of the testing data by the synthetic data. The coverage values are multiplied by 100 in the table. Clearly TabDDPM

has the best performance. The results also reveal that all other models suffer mode dropping or collapse [26], as shown by their small coverage values.

Table 5: The 5% quantile of the distances to the nearest neighbor.

model	dcr rs	dcr hs	rs/hs	dcr rr	dcr ss	percentile
GaussianCopula	**0.118 (0.024)**	0.118 (0.024)	1.001 (0.005)	0.003 (0.000)	0.069 (0.005)	5
CTGAN	0.012 (0.002)	0.012 (0.002)	1.016 (0.009)	0.004 (0.001)	0.014 (0.002)	5
TVAE	0.007 (0.000)	0.007 (0.000)	1.002 (0.010)	0.003 (0.000)	0.006 (0.000)	5
CTABGAN	0.027 (0.003)	0.027 (0.002)	1.004 (0.004)	0.004 (0.001)	0.029 (0.003)	5
stasy	0.026 (0.001)	0.026 (0.001)	1.010 (0.005)	0.003 (0.000)	0.020 (0.001)	5
TabDDPM	0.003 (0.000)	0.003 (0.000)	1.006 (0.011)	0.003 (0.000)	0.003 (0.000)	5

Privacy Leakage Metric We report the privacy leakage metric results in Table 5, where "dcr rs" is the distance to the closest synthetic record from each real training data, "dcr hs" is the DCR from holdout data to synthetic data, "dcr _rr" the DCR within real data, "dcr ss" is the DCR within synthetic data, "rs/hs" is the ratio $\frac{d_\alpha(r,s)}{d_\alpha(h,s)}$, and "percentile" is the α of the DCRs, as described in Section 4.2.

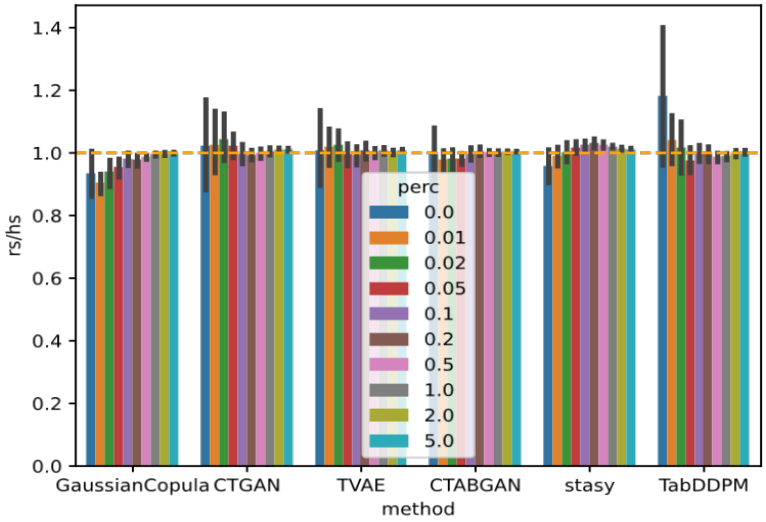

Figure 1: Privacy leakage assessment

Based solely on "dcr rs", as commonly used in previous literature [28], the Gaussian Copula model has the best privacy preservation, while the TabDDPM model has the worst. However, the results also demonstrate that none of the "dcr _rs" is smaller than "dcr rr", which implies that actually there is possibly no privacy leakage as the synthetic data is far away from the real data. Just as we discussed in Section 4.2, "dcr rs" alone may be insufficient to assess the risk of privacy leakage.

With the proposed DCR ratio metric, we calculate the *rDCR* for different percentile α values, and present the results in Figure 1. As shown in the figure, contrary to the above conclusion, the Gaussian Copula model is vulnerable to membership inference attacks at very small values of α,

where its DCR ratio is smaller than 1, while all the other models appear to be robust against distance-based membership inference attacks, as their DCR ratio remains approximately 1. This finding highlights the advantage of the percentile-based ratio metric.

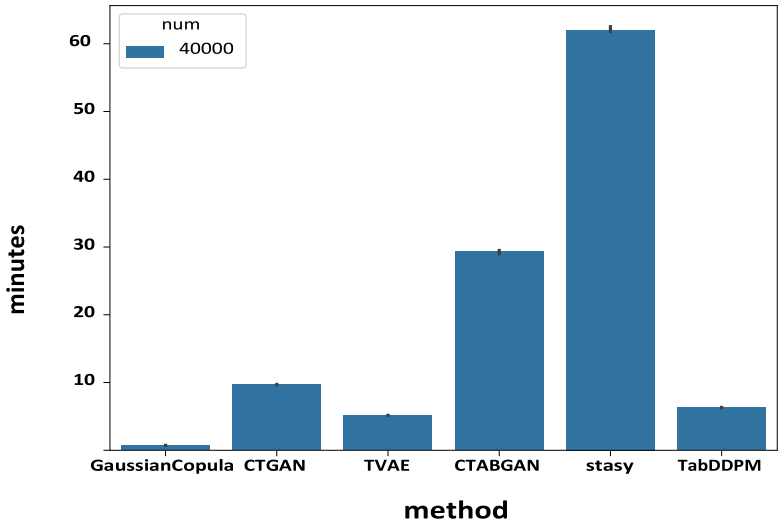

Figure 2: Complexity based on DCR ratio (running time in minutes).

Complexity We evaluate the complexity of these models by comparing their running times. Note that the running time for diffusion-based models includes both the training and sampling time, whereas for other models, it consists only of the training time. The results are presented in Figure 2. The running times are reported in minutes, obtained from a machine with Intel(R) Core(TM) i99900X CPU, 64G memory, and GeForce RTX 2080. The results demonstrate that CTABGAN and STaSy models have much higher time complexity than others. Although the Gaussian Copula model has the fast training speed, its performance is not satisfactory, especially as evidenced by its minimum coverage values, severe mode collapse and possible privacy leakage. The results indicate that TabDDPM achieves the best balance between speed and performance.

CHAPTER 6

Conclusions

In this project, we conduct a systematic evaluation of generative models for synthetic tabular transportation data generation. The evaluation is conducted based on a variety of metrics, including downstream tasks performance, distribution similarity, generation diversity, and privacy leakage. We also evaluate these models on our novel graph similarity and DCR ratio metrics.

The results indicate that TabDDPM achieves the overall best performance across various metrics. However, it appears that TabDDPM may struggle to handle categorical variables with hundreds of classes. Additionally, the findings reveal the performance gap of the current generative models and the prevalence of mode collapse, underscoring the need to develop models specifically tailored to domains such as transportation.

Furthermore, extending the evaluation beyond the New York City taxi data is expected to offer more insights on the current tabular generative models.

REFERENCES

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
2. ATOCKAR: Riding with the stars: Passenger privacy in the nyc taxicab dataset. https://agkn.wordpress.com/2014/09/15/ riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/
   (2014), [Online; accessed 2024-11]
3. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
4. Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE transactions on neural networks and learning systems (2022)
5. Boudewijn, A.T.P., Ferraris, A.F., Panfilo, D., Cocca, V., Zinutti, S., De Schepper, K., Chauvenet, C.R.: Privacy measurements in tabular synthetic data: State of the art and future research directions. In: NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI (2023)
6. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 1897–1914. IEEE (2022)
7. Correa, D.: Exploring the taxi and uber demands in new york city: An empirical analysis and spatial modeling. Available at SSRN 4229042 (2017)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Hilprecht, B., Ha¨rterich, M., Bernau, D.: Reconstruction and membership inference attacks against generative models. arXiv preprint arXiv:1906.03006 (2019)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Hoogeboom, E., Nielsen, D., Jaini, P., Forr´e, P., Welling, M.: Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems **34**, 12454–12465 (2021)
12. Kim, J., Lee, C., Park, N.: Stasy: Score-based tabular data synthesis. arXiv preprint arXiv:2210.04018 (2022)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Kotelnikov, A., Baranchuk, D., Rubachev, I., Babenko, A.: Tabddpm: Modelling tabular data with diffusion models. In: International Conference on Machine Learning. pp. 17564–17579. PMLR (2023)
15. scikit learn: Coefficient of determination. https://scikit-learn.org/stable/ modules/generated/sklearn.ensemble.GradientBoostingRegressor.html# sklearn.ensemble.GradientBoostingRegressor.score (2024), [Online; accessed 2024-11]
16. scikit learn: Gradient boosting regressor. https://scikit-learn.org/stable/ modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
    (2024), [Online; accessed 2024-11]

17. Mauro, G., Luca, M., Longa, A., Lepri, B., Pappalardo, L.: Generating mobility networks with generative adversarial networks. EPJ data science **11**(1), 58 (2022)
18. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning. pp. 7176–7185. PMLR (2020)
19. Patki, N., Wedge, R., Veeramachaneni, K.: The synthetic data vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 399–410 (2016). https://doi.org/10.1109/DSAA.2016.49
20. Platzer, M., Reutterer, T.: Holdout-based empirical assessment of mixed-type synthetic data. Frontiers in big Data **4**, 679939 (2021)
21. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 3–18 (2017). https://doi.org/10.1109/SP.2017.41
22. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Scorebased generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
23. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data–anonymisation groundhog day. In: 31st USENIX Security Symposium (USENIX Security 22). pp. 1451–1468 (2022)
24. Taxi, Commission, L.: Tlc trip record data. https://www.nyc.gov/site/tlc/ about/tlc-trip-record-data.page (2023), [Online; accessed 2024-11]
25. Taxi, Commission, L.: 2015 green taxi trip data. https://data.cityofnewyork.us/Transportation/2015-Green-Taxi-Trip-Data/gi8d-wdg5/about_data (2024), [Online; accessed 2024-11]
26. Thompson, R., Knyazev, B., Ghalebi, E., Kim, J., Taylor, G.W.: On evaluation metrics for graph generative models. arXiv preprint arXiv:2201.09871 (2022)
27. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Advances in neural information processing systems **32** (2019)
28. Zhao, Z., Kunar, A., Birke, R., Chen, L.Y.: Ctab-gan: Effective table data synthesizing. In: Asian Conference on Machine Learning. pp. 97–112. PMLR (2021)