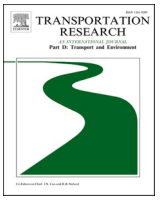
Transportation Research Part D 127 (2024) 104067

Contents lists available at ScienceDirect

Transportation Research Part D

journal homepage: www.elsevier.com/locate/trd



Fake it till you make it: Synthetic data for emerging carsharing programs

Tobias Albrecht [a,b,c,*], Robert Keller [a,b,d], Dominik Rebholz [a,b,d], Maximilian Röglinger [a,b,c]

[a] Branch Business & Information Systems Engineering of the Fraunhofer FIT, Bayreuth, Germany
[b] FIM Research Center for Information Management, Bayreuth, Germany
[c] University of Bayreuth, Bayreuth, Germany
[d] Kempten University of Applied Sciences, Kempten, Germany

ARTICLE INFO

Keywords:
Carsharing
Electric vehicle
Generative adversarial network
Machine learning
Shared mobility
Synthetic data

ABSTRACT

Carsharing is an integral part of the transformation toward flexible and sustainable mobility. New carsharing programs are entering the market to challenge large operators by offering innovative services. This study investigates the use of generative machine learning models for creating synthetic data to support carsharing decision–making when data access is limited. To this end, it explores the evaluation, selection, and implementation of leading-edge methods, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), to generate synthetic tabular transaction data of carsharing trips. The study analyzes usage data of an emerging carsharing program that is expanding its services to include free-floating electric vehicles (EVs). The results show that augmenting real training data with synthetic samples improves predictive modeling of upcoming trips by up to 4.63%. These results support carsharing researchers and practitioners in generating and leveraging synthetic mobility data to develop solutions to real-world decision support problems in carsharing.

1. Introduction

Carsharing has gained traction as a sustainable concept to address the immediate challenges of urban mobility (Ferrero et al., 2018; Hu et al., 2018a). It has been found to reduce CO_2 emissions, noise pollution, congestion, and parking shortages by decreasing car ownership or postponing vehicle purchase decisions (Nijland and van Meerkerk, 2017; Kim et al., 2019; Vélez, 2023), lowering fuel consumption and annual vehicle kilometers traveled (Chen and Kockelman, 2016; Meng et al., 2020), and fostering intermodal transportation (Amatuni et al., 2020; Chicco and Diana, 2021). As a prominent example of the sharing economy, carsharing induces changes in the mobility behavior of its users toward short-term vehicle access as a service that can complement public transport and active travel (Münzel et al., 2018; Vanheusden et al., 2022). Further, carsharing can accelerate the transition to green transport through the increasing integration of electric vehicles (EVs) in modern carsharing fleets (Hu et al., 2019; Luna et al., 2020; Prinz et al., 2020a; Shaheen et al., 2020; He and Chen, 2021; Hoerler et al., 2021). Fueled by these benefits, the demand for carsharing services is projected to continue to grow (Mounce and Nelson, 2019; Shaheen and Cohen, 2020). While carsharing programs were previously the preserve of large operators from the private sector, public institutions and small operators are increasingly entering the market (Zhang

* Corresponding author at: Fraunhofer FIT, Wittelsbacherring 10, 95444 Bayreuth, Germany.
E-mail address: tobias.albrecht@fit.fraunhofer.de (T. Albrecht).

https://doi.org/10.1016/j.trd.2024.104067
Received 7 August 2023; Received in revised form 22 December 2023; Accepted 5 January 2024
Available online 17 January 2024

et al., 2020; Baumgarte et al., 2022). In the course of this development, emerging programs are coming up with innovative service offers and tapping into new markets that are no longer limited to metropolitan areas but increasingly include rural and suburban areas (Rotaris and Danielis, 2018; Lagadic et al., 2019; Illgen and Höck, 2020).

To succeed in this competitive market, carsharing operators and policy makers are required to continuously realign service offers and policies with user needs (Hu et al., 2018a; Münzel et al., 2019). Consequently, the introduction of new carsharing programs as well as the implementation of innovative service offers involve a variety of operational and strategic decisions that benefit from understanding and anticipating user behavior (Golalikhani et al., 2021a). Prescriptive research aims to provide data-driven decision support in areas related to the business model and operating area (Perboli et al., 2018; Hahn et al., 2020), the location, number, and capacity of stations and charging possibilities (Zhang et al., 2020; Abbasi et al., 2021; Ma and Xie, 2021), the ideal fleet size and distribution of vehicles (Hu and Liu, 2016; Jian et al., 2017; Illgen and Höck, 2019; Brendel et al., 2022), user demand (Jorge and Correia, 2013; Yoon et al., 2019; Cheng et al., 2021), and pricing schemes (Yang et al., 2022; Brendel et al., 2023). As the number of carsharing users and the demand for more flexible service offers grow, decision-making for carsharing operators becomes ever more complex (Laporte et al., 2018). Hence, insights and predictions of user behavior and trip characteristics from machine learning (ML) models have become indispensable for carsharing operators to make informed decisions and optimize their systems. Previous literature highlights the high value of ML for predicting, inter alia, trip distances (Baumgarte et al., 2022), user demand (Cocca et al., 2020; Cheng et al., 2021; Prinz et al., 2022), and supply imbalances (Willing et al., 2017; Wang et al., 2021). Consequently, large carsharing operators increasingly leverage their rich database to generate insights for improving the efficiency and sustainability of their operations, reduce costs, and provide a better service experience for their users (Golalikhani et al., 2021a; Yao et al., 2022).

Emerging carsharing programs are mostly excluded from this frontline industrial development. While ML models promise high prediction accuracy and flexible adaptation to a variety of decision support problems, their practical value is also highly dependent on their deployment and on the available input data (Albrecht et al., 2021; Shrestha et al., 2021). In addition to budget constraints and lower technological expertise, many small carsharing operators struggle with limited data availability and quality (Lagadic et al., 2019). Further, municipal operators in particular must comply with strict data protection regulations (Vanheusden et al., 2022). In this connection, synthetic data created by generative ML models such as generative adversarial networks (GANs) and variational autoencoders (VAEs) present a promising path to overcome the challenge of restricted access to high-quality and privacy-preserving input for the training of data-intensive ML models.

In transportation research, generative ML has lately been successfully employed for population synthesis and activity scheduling (Garrido et al., 2020; Kim et al., 2022), urban vehicle trajectory modeling (Zhang et al., 2019; Choi et al., 2021), and learning probabilistic dependencies in travel behavior data (Wong and Farooq, 2020). First steps have also been made to investigate the use of synthetic training samples for prediction tasks like travel mode detection (Li et al., 2020) and road traffic forecasting (Boquet et al., 2020). While these studies offer valuable insights on traffic management and public planning, an in-depth understanding of the use of generative ML and synthetic data for innovative mobility options like carsharing is needed to foster their development.

Against this backdrop, this study aims to explore the use of synthetic data to support carsharing decision–making by overcoming the barrier of limited data access during the introduction and expansion of new services. To this end, it means to investigate the evaluation, selection, and implementation of generative ML models to create synthetic tabular transaction data of carsharing trips for more accurate predictions of user behavior when data access is limited. Thus, this study raises the following research questions: *Can synthetic data created by generative ML models support decision-making in carsharing? And what are appropriate methods?*

To pursue this research objective, we employ a systematic ML workflow (Kühl et al., 2021; Shrestha et al., 2021) as a rigorous approach to the generation of synthetic data as well as to the evaluation of their fidelity and their utility in the context of real-world prediction tasks. Consistent with the application–oriented nature of our study, we consider the case of an emerging carsharing program that is expanding its services to include free-floating EVs and aims to obtain more reliable predictions of trip distances and usage times during the service's ramp-up. In this connection, we analyze how well synthetic data generated by GANs, VAEs, and benchmark models serve the purpose of enhancing the available database to improve the training and performance of prediction models. We investigate the use of synthetic data for replacing, rebalancing, and augmenting real training data drawing on two evaluation protocols that correspond to the prediction of two different target variables.

The findings of this study contribute to research by presenting novel insights on the use of generative ML for the creation of synthetic mobility data. They help understand the methods and approaches for generating and leveraging synthetic tabular transaction data of carsharing trips for more accurate predictions of user behavior. In this way, the results advance current studies concerned with developing methodological solutions to the prevalent decision support problems in carsharing (Golalikhani et al., 2021a) and particularly add to the body of knowledge on small, municipal, and emerging carsharing programs as well as on the introduction and expansion of EV carsharing services (Xu and Meng, 2019; Luna et al., 2020). In addition to its theoretical contribution, our study supports practitioners with limited data access in using generative ML models to enhance their available database, achieve more accurate predictions of user behavior, and make more informed operational and strategic decisions when launching or expanding carsharing services (e.g., developing user incentives or dynamic pricing models).

The remainder of this paper is structured as follows. In Section 2, we review relevant literature and provide theoretical background on carsharing decision support as well as on generative ML and synthetic data. We present our research design in Section 3. In Section 4, we illustrate the results of our study before discussing them with a focus on their theoretical and managerial implications in Section 5. Section 6 concludes by pointing out limitations and indicating avenues for future research.

2. Related work

2.1. Carsharing decision support

Carsharing programs rely on platform mediation to match provider resources and user mobility demands (Nansubuga and Kowalkowski, 2021). As such, they form an integral part of the ongoing shift from vehicle ownership to vehicle access as a service (Cohen and Kietzmann, 2014; Schmöller and Bogenberger, 2020). Carsharing provides registered users with short-term access to a fleet of shared vehicles distributed within a designated operating area and offers a pay-per-use model, usually based on the trip distance and the usage time (Shaheen et al., 2006). In general, carsharing programs can be classified into three main types: one-way, roundtrip, and free-floating carsharing (Ferrero et al., 2018; SAE International, 2021). One-way and roundtrip services are station–based (i.e., vehicle pick-up and drop–off are limited to stations provided by the operator). While one-way services allow users to end their trips at any given station, roundtrip services require users to return the vehicle to the same station they started from. Free–floating services provide more flexibility by allowing the users to start and end their trips anywhere within the operating area (Nourinejad and Roorda, 2015; Münzel et al., 2019). The added convenience of free–floating trips has led to the growing popularity of this type of service (Vanheusden et al., 2022), yet this trend also poses new challenges such as refueling (Meng et al., 2020), maintenance trips (Giordano et al., 2021), or vehicle relocation to counter supply and demand imbalances (Weikl and Bogenberger, 2015; Xu and Meng, 2019) which increase the overall decision-making complexity for operators.

Moreover, carsharing programs can be categorized according to the type of engine used in their vehicles (i.e., electric or internal combustion engine). In recent years, improved technology and favorable policy have led to the rise of EV-based carsharing schemes (Shaheen et al., 2020). Encouraged by the positive impact of EV carsharing on the urban environment (Luna et al., 2020; Li et al., 2021), public institutions aim to promote sustainable mobility by providing incentives for EV fleets (e.g., through tax reliefs, expansion of charging infrastructure, and access to bus lanes) (Xu et al., 2021) and operators strive to attract more environmentally aware users with EVs (Brendel et al., 2018; Giordano et al., 2021). In addition to the environmental benefits of conventional carsharing, EV carsharing further reduces CO_2 emissions (Wappelhorst et al., 2014; Boyacı and Zografos, 2019) and has been shown to facilitate mass adoption of EVs by providing a positive driving experience and reducing potential skepticism such as range anxiety (Brendel et al., 2018). Besides, the higher annual mileage of carsharing vehicles results in a shorter payback period of acquisition costs for operators and a more positive life cycle assessment (LCA) compared to private EVs (Burghard and Dütschke, 2019). In order to fully exploit the benefits of EVs, carsharing providers must also deal with new managerial challenges like the limited range of EVs per charging cycle and the associated need for frequent recharging (Hu et al., 2019; Cui et al., 2022; Yao et al., 2022). This entails new decision-making areas for operators such as the optimal location of charging points (Cocca et al., 2019; Lai et al., 2022) and the harmonization of EV charging with the driving habits of the users (Boyacı and Zografos, 2019; Shen et al., 2019).

In addition to diversifying their services, carsharing operators are also beginning to expand to non-urban areas to attract new user segments beyond the saturated urban market (Wappelhorst et al., 2014; Illgen and Höck, 2020). While less developed public transport networks and the existing demand for individual and first-mile/last-mile mobility favor carsharing in rural and small urban areas, low population density as well as a higher car ownership level and different usage habits (e.g., infrequent but longer non-commuting trips) pose new challenges to carsharing programs (Rotaris and Danielis, 2018; Lagadic et al., 2019). To deal with these circumstances, carsharing operators are adapting their business models to be more socially oriented and to involve local authorities and municipalities (e.g., through subsidized pricing and cooperation with local public transport operators). Additionally, the number of carsharing programs operated directly by municipalities is growing (Rotaris and Danielis, 2018; Baumgarte et al., 2022).

To thrive with new services and in new markets, both emerging and established carsharing operators need to focus on improving supply attributes related to vehicle availability and user convenience, as they directly influence perceived service quality and usage intention (Niels and Bogenberger, 2017; Hu et al., 2018b; Golalikhani et al., 2021b). This involves optimizing the number, types, and (re-)location of vehicles (Hu and Liu, 2016; Jian et al., 2017; Illgen and Höck, 2019; Prinz et al., 2020), charging or refueling policies (Weikl and Bogenberger, 2015; Meng et al., 2020), reservation and pricing (Molnar and Correia, 2019; Yang et al., 2022) as well as the location, size, and quantity of stations for station-based services (Correia and Antunes, 2012; Zhang et al., 2020; Abbasi et al., 2021). These decisions are often based on a thorough understanding of factors for carsharing adoption and demand, user groups, and travel behavior from descriptive research (Costain et al., 2012; de Lorimier and El-Geneidy, 2013; Hu et al., 2018a; Baumgarte et al., 2021) as well as on precise predictions of user demand and upcoming trip characteristics (Lei et al., 2020; Wang et al., 2021). Regarding the latter, operators and policy makers increasingly rely on state-of-the-art ML models for data-driven decision support (Willing et al., 2017; Prinz et al., 2022).

Previous research on carsharing demand modeling draws on various forms of ML like gradient boosting machines (GBM) to predict station-level vehicle demand for one-way carsharing programs (Wang et al., 2021) and to develop a spatial decision support system for demand imbalances in free–floating carsharing (Willing et al., 2017), recurrent neural networks (RNN) to optimize the proactive relocation policy for on-demand mobility services (Lei et al., 2020) or long short-term memory networks (LSTM) to predict service demand in free-floating carsharing (Cocca et al., 2020; Alencar et al., 2021a) and one-way station-based carsharing programs (Wang et al., 2020; Brahimi et al., 2022). Other studies aim to optimize vehicle relocation and positioning for free-floating services by predicting the time to pick-up (Kostic et al., 2021) and by designing a competitor-aware vehicle positioning model (Schroer et al., 2022) based on deep feedforward neural networks (DNN). In this connection, convolutional neural networks (CNN) are employed by Zhu et al. (2019) for carsharing flow prediction and by Chang et al. (2022) for optimizing vehicle relocations and staff movements among different carsharing providers of one-way station-based programs. Ren et al. (2020) develop a station scheduling method for vehicle rebalancing based on reinforcement learning while Prinz et al. (2022) present vehicle relocation strategies for free-floating

services based on demand predictions using random forests (RF). To optimize the location of carsharing stations based on user demand, Zhu et al. (2017) developed a framework relying on edge computing and a stacked autoencoder. In this context, Ma et al. (2022) use CatBoost and shapley additive explanations (SHAP) to predict development patterns of one-way carsharing stations considering the occupancy rate and built environment. Baumgarte et al. (2022) and Cheng et al. (2021) employ gradient boosting approaches and SHAP to reveal the factors influencing carsharing trip distance and booking demand in station-based carsharing.

The emergence of big urban mobility data has resulted in new opportunities for the application of ML and predictive analytics in carsharing (Schroer et al., 2022). Previous work draws on large data sets with at least several hundred thousand trips that they acquired through direct cooperation with established carsharing providers, via application programming interfaces (APIs) to the digital platforms of large carsharing companies that allow for automated and real-time data retrieval (car2go, 2019), or using data sharing standards (Open Mobility Foundation, 2023) and open data initiatives (Ciociola et al., 2017; Alencar et al., 2021b; Schroer et al., 2022). While these trends are expected to continuously improve the real–time carsharing data availability, small and emerging carsharing programs are mostly excluded from this development. Due to an initially small user base, limited financial and human resources as well as a lack of standardized systems and technological capabilities, emerging carsharing programs often face the challenge of restricted access to representative data of their services (Brendel et al., 2017; Lagadic et al., 2019). Limited expertise in dealing with data privacy restrictions and establishing data sharing policies often adds to these barriers. This is even more applicable to municipal carsharing programs that are increasingly being established in non-urban areas (Rotaris and Danielis, 2018; Vanheusden et al., 2022). Consequently, data acquisition and quality are still major challenges for many carsharing operators on the market, preventing them from keeping up with the latest technological developments and sophisticated business models of large established programs. However, these challenges are not yet sufficiently addressed by previous research.

2.2. Synthetic data and generative machine learning

In real-world settings, data access is often limited by privacy restrictions (e.g., disclosure of sensitive information) (Shrestha et al., 2021), prohibitively expensive or time-consuming data collection (e.g., data labeling) (Zhou et al., 2017), or lack of data quality and representativeness (e.g., noise or class imbalance) (Gudivada et al., 2017). Synthetic data generation has emerged as a valuable technique to overcome these challenges. In contrast to data collected from real-world sources, synthetic data refers to data that is artificially generated by purpose-built models (Jordon et al., 2022). Going back to the early work on statistical disclosure control by Rubin (1993) and Little (1993), synthetic data has recently come to the fore as high-quality, privacy-preserving input for the training of data–intensive ML models (Nikolenko, 2021; Figueira and Vaz, 2022). The results of such ML models scale with the quality and quantity of training data (Sengupta et al., 2020). In this connection, the application potential of synthetic data includes replacing or augmenting training data when real-world data is unavailable, sensitive to data protection, noisy, or imbalanced (Tanaka and Aranha, 2019; Carvajal-Patiño and Ramos-Pollán, 2022).

When real data is scarce, augmenting training data with high-quality synthetic data can help make ML models more robust and less prone to overfitting by introducing new additional data points as variations or perturbations to the existing data (Nikolenko, 2021; Figueira and Vaz, 2022). As such, synthetic data can also complement underrepresented regions of the data space and mitigate data imbalances, providing a more comprehensive representation of the underlying patterns and structures in the data (Jordon et al., 2022). In contrast, simply duplicating the available training data creates identical samples that do not provide new information to the model, potentially leading to overfitting to duplicated instances, lack of data diversity, and the amplification of biases in the data (Zhou et al., 2017). In addition, synthetic data can help develop and test models in a controlled environment (e.g., by introducing predefined distributions or characteristics to the data) (Bolón-Canedo et al., 2013) as well as ensure data privacy and enable data sharing (e.g., by controlling information release) (Snoke et al., 2018; Jordon et al., 2020).

Synthetic data is created by generative models that are characterized by their ability to synthesize new data by learning a distribution $p_{model}(x)$ from training samples x that approximates $p_{data}(x)$ as closely as possible (Goodfellow et al., 2020). Generally, generative models are classified into two main types: explicit density models and implicit density models (Goodfellow, 2017). The former make explicit probabilistic assumptions of the data distribution in the form of a probability density function $p_{model}(x;\theta)$. This density may be computationally tractable (e.g., deep belief networks (DBNs)) where the log–likelihood of training data can be maximized directly or intractable, meaning that either variational approximations (e.g., VAEs) or Monte Carlo approximations (e.g., restricted Boltzmann machines (RBMs)) are needed to maximize the likelihood. In contrast, implicit density models are trained without explicitly defining a density function of the data space. Instead, they interact with the distribution by learning only a tractable sample

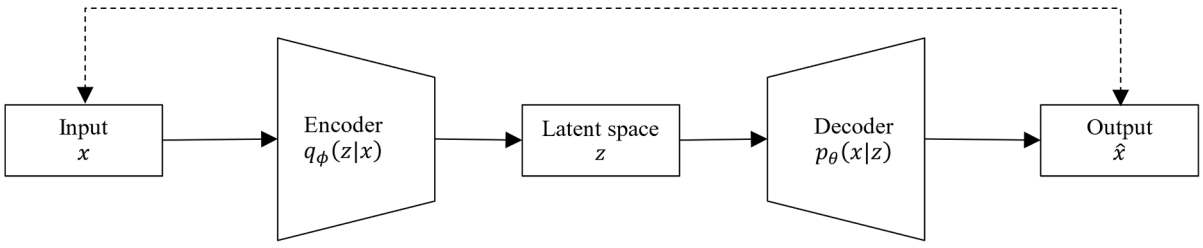


Fig. 1. Schematic VAE architecture.

generation process that may be incremental based on Markov chains (e.g., generative stochastic networks (GSNs)) or in a single generation step (e.g., GANs). Broadly, implicit density models define a way to stochastically transform an existing sample to obtain new samples from the same distribution (Goodfellow, 2017). Today, the most widely used approaches for generative modeling are VAEs and GANs due to their high performance in a variety of application areas with multimodal data such as images (Alqahtani et al., 2021), process logs (van Dun et al., 2022), and natural language (Hsu et al., 2017).

VAEs introduced by Kingma and Welling (2013) are currently gaining traction as one of the most popular models for synthetic data generation. Resembling standard autoencoders, VAEs are unsupervised learning algorithms consisting of two main components (Fig. 1): (1) an encoder network $q_\phi(z|x)$ with parameters ϕ that maps the input data x into a lower-dimensional latent space z with a prior distribution $p(z)$ and (2) a decoder network $p_\theta(x|z)$ with parameters θ that reconstructs the latent code as output $\hat{x}$ to match the input data (Kingma and Welling, 2019). Generally, the distribution of the real data mapped to the latent space of a standard autoencoder is sparse. To enable sampling new data, VAEs force a continuous latent space by learning the parameters of a probability distribution representing the original data as vectors of means and standard deviations (i.e., μ and σ). New samples are then generated by sampling a point from this regularized latent space and passing it through the decoder. The loss function that is minimized during VAE training is then composed of the reconstruction error (i.e., the mean squared loss of the input and reconstructed output) and the Kullback–Leibler (KL) divergence (i.e., the divergence between the encoder distribution $q_\phi(z|x)$ and $p(z)$) as a regularization term for the latent space. It can then be written as $L(x,\hat{x}) + D_{KL}\left(q_\phi(z|x) \parallel p(z)\right)$. VAEs have successfully been applied for synthetic data generation in different fields such as crash data analysis (Islam et al., 2021), construction management (Davila Delgado and Oyedele, 2021), and breast cancer diagnosis (Inan et al., 2023).

Introduced by Goodfellow et al. (2014), GANs are based on the idea of game theory. Their architecture consists of two neural networks: a generator G and a discriminator D (Fig. 2). The generator takes random noise z sampled from a prior distribution $p_z(z)$ as input to generate new data $G(z)$. The discriminator then aims to distinguish between real data samples x and synthetic samples $G(z)$. Thus, the objective of D is to correctly classify the data source with high accuracy, while G aims for an equal performance of D for x and $G(z)$ (Wang et al., 2017). Formally, this results in a two–player minimax game with the value function $V(D,G)$:

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

GANs derive their name from the adversarial optimization process with opposing goals during model training, where the generators' and discriminators' weights are alternately updated through backpropagation from their respective loss functions V until the Nash equilibrium (i.e., $D(x) = \frac{1}{2}$ for all x) is reached (Ratliff et al., 2013). Based on the original architecture, numerous GAN variants have been proposed to address challenges in the learning process of the initial GAN algorithm. Most notably, CGAN introduces a conditional setting to direct the data generation process based on class labels (Mirza and Osindero, 2014), InfoGAN adds control variables to the learning process in an unsupervised manner (Chen et al., 2016), and WGAN addresses instability problems like mode collapse during model training (Arjovsky et al., 2017). While GANs have primarily been used in computer vision (e.g., for image generation, face synthesis, and image translation) (Alqahtani et al., 2021), their ability to learn probability distributions and draw high-quality realistic samples has encouraged the recent development of GANs for tabular data that requires modeling complex distributions of diverse data types (Lei and Veeramachaneni, 2018; Park et al., 2018; Lei et al., 2019).

The performance of generative ML models with respect to the quality of the created synthetic data sets can be evaluated in terms of privacy, fidelity, and utility (Jordon et al., 2022). Privacy measures notwithstanding, this task is equivalent to assessing the dissimilarity between the distributions $p_{data}(x)$ and $p_{model}(x)$ (Borji, 2022). In this connection, fidelity refers to the extent to which the synthetic data accurately represents the characteristics and statistical properties of the real data. Typically, evaluating fidelity involves distance measures and visualizations to compare the distributions of real and synthetic data (Figueira and Vaz, 2022). The utility of a synthetic data set, on the other hand, refers to the usefulness or effectiveness of the synthetic data for a specific purpose or task. It assesses the extent to which the synthetic data can replicate, replace, or augment the real data for data-driven tasks (Snoke et al., 2018; Jordon et al., 2022). For instance, it evaluates whether the generated data can capture the underlying patterns and structures of the original data in a way that allows ML models trained on it to perform well on unseen real data (i.e., ML efficacy).

In transportation research, generative ML approaches are on the rise but are still dominated by simulation-based approaches and

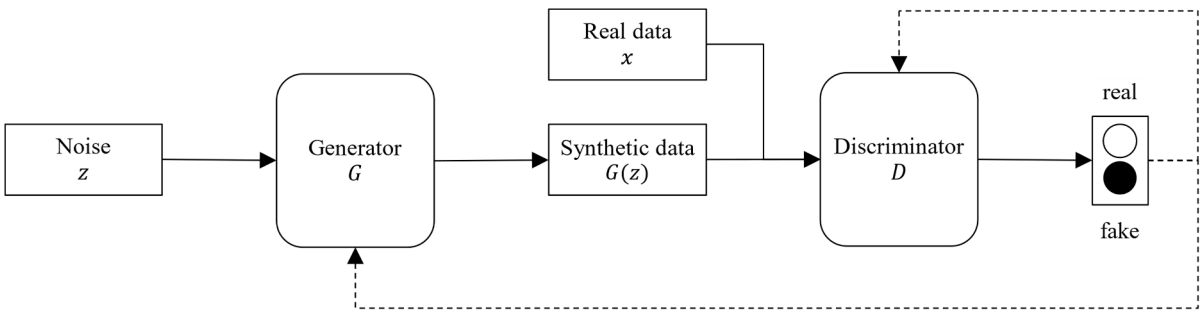


Fig. 2. Schematic GAN architecture.

standard statistical models for synthetic data generation (Wong and Farooq, 2020). Previous studies in this area mainly focus on model development for domain-specific applications such as population synthesis for agent-based transport models and the generation of individual travel behavior data for trip forecasting. For instance, Etxandi-Santolaya et al. (2023) estimate EV battery capacity requirements based on synthetic driving cycles created via stochastic simulation from the empiric probability functions of real data. Brendel et al. (2017) generate rental data for carsharing relocation simulations using a combination of decision trees and a Gaussian mixture model (GMM) while Wong and Farooq (2020) propose an RBM-based model for analyzing and simulating travel behavior data for demand analysis. VAEs are applied by Islam et al. (2021) to augment roadway network data for crash prediction and by Yao and Bekhor (2022) for route choice set generation. Further, Boquet et al. (2020) draw on VAEs to generate and impute traffic data as input for road traffic forecasting systems and use their latent space for model selection and traffic anomaly detection. Garrido et al. (2020) apply VAEs and GANs to travel survey data for population synthesis as input for agent-based systems for long-term travel forecasts. GANs are also deployed by Li et al. (2020) who create synthetic GPS data to improve the accuracy of travel mode detection models. Choi et al. (2021) and Zhang et al. (2019) propose GANs for generating urban vehicle trajectory data as well as for learning its travel time distribution. In addition, Kim et al. (2022) find that conditional GANs outperform other models when imputing the socio-demographic attributes and trip purpose of transit passengers' mobile data.

Previous studies highlight the great potential of generative ML in transportation management and planning. However, most studies primarily focus on tailored methodological solutions for simulation tasks such as trajectory modeling and population synthesis, while research has just begun to investigate the use of synthetic training samples for prediction tasks. Thus, our work contributes to this body of knowledge by being the first to employ synthetic data created by generative ML models to support the prediction of user behavior during the introduction and expansion of innovative mobility options.

3. Research design

This study investigates approaches to effectively generate and leverage synthetic trip data in carsharing to mitigate the challenge of limited data access during the introduction and expansion of new services and enable more accurate predictions of upcoming trip characteristics. To this end, it analyzes the fidelity and utility of synthetic data in the context of real-world prediction tasks. This research examines the case of an emerging municipal carsharing program that is expanding its services to include free-floating EVs but lacks the large database necessary to make reliable predictions about user behavior and trip characteristics. To rigorously pursue our research goal, we draw on a four–phase research approach following a systematic ML workflow (Kühl et al., 2021; Shrestha et al., 2021) that reflects the application–oriented nature of our study (Fig. 3).

First, we obtained a thorough *data and business understanding*. For this purpose, we worked closely with the carsharing operator to understand its operational processes and business model. We evaluated the available database, reviewed related documents, and had several discussions with the operator to elaborate the status quo and the underlying problem statement from both a business and data perspective. The present carsharing program is located in a medium-sized German town with a population of just under 300,000. It is fully owned by the city and operated by its public utility. Compared to large-scale private sector programs, the municipal nature of the present carsharing program involves several particularities that affect its management and expansion. For instance, social and

Iteration	Research phase	Research steps	
1	Data and business understanding	• **Business analysis**	Discussions with the operator, review of relevant documentation, and derivation of the problem statement
		• **Data analysis**	Descriptive data analysis and quality check
2	Data preparation	• **Data merging**	Aggregation of transaction data, user data, and weather data
		• **Feature engineering**	Geocoding and creation of variables related to prior usage
		• **Data preprocessing**	Feature selection, outlier detection, exclusion of operator trips
		• **Data splitting**	80/20 split in training and test data
3	Data generation	• **Model creation**	CTGAN, TVAE, GC, and SMOTE -NC
		• **Model training and optimization**	5-fold cross validation, random grid search
		• **Synthetic data generation**	Synthetic data sets for replacing, rebalancing, and augmenting the real training data
4	Evaluation and deployment	• **Fidelity evaluation**	Statistical and graphical evaluation of data fidelity
		• **Utility evaluation**	Predictive evaluation of data utility (i.e., prediction of usage times and trip distances using DNN, RF, and XGBoost)
		• **Model selection**	Selection of the best performing generative model
		• **Deployment**	Derivation of implications for model deployment

Fig. 3. Iterative research process.

environmental responsibility is an integral part of its business model, and data processing and transfer are subject to strict data protection regulations.

Established in 2015 as a small roundtrip station-based program, it has since increased its fleet to 300 vehicles and 110 stations. To encourage local sustainable mobility and to reach new user groups, the operator now aims to expand its services to include free-floating EVs. Following a successful trial period, the new EV fleet currently consists of 13 vehicles that are distributed throughout the downtown area and can be reserved or accessed directly via a mobile app. Without having to specify the duration of usage beforehand, users can end their trip by parking the EV in any public parking space in the operating area. The usage fee is then calculated based on the usage time and the distance traveled. The basic pricing plan further includes a one-time subscription fee and a monthly membership fee. To optimize EV availability and the overall user experience, the operator aims to obtain reliable predictions of upcoming trip characteristics (i.e., trip distance and usage time). In this way, the program seeks to establish more efficient charging processes (i.e., provide users with real–time suggestions of charging stations and times based on the current battery level of the EV and predictions of the trip distance) and to improve its booking system (i.e., provide users with vehicle availability forecasts despite the open-ended nature of EV bookings). However, data availability is limited due to the short period of service and the restricted user base in a small urban area. Hence, the initially available data includes a transaction data set of 6,479 trips made with free–floating EVs between January and December 2021, a user data set with demographic and contract information of 771 active users who made at least one trip with an EV during the period of analysis, and hourly location-specific weather data for the analyzed period (OpenWeather, 2023).

Second, during *data preparation*, we merge the transaction data, user data, and weather data to obtain one aggregated data set (Schmöller et al., 2015). In the course of feature engineering, we created new variables indicating the user-specific prior usage behavior (e.g., prior distance traveled), merged categorical values where appropriate, and geocoded the start and end locations of each trip (Wielinski et al., 2019). We further prepared the data by excluding maintenance and relocation trips, employing correlation-based feature selection and interquartile range for outlier detection (Kuhn and Johnson, 2019). The resulting data set includes 5,945 trips described by 20 explanatory variables from six categories (i.e., time-related, location-related, car-related, user-related, usage-related, and weather-related variables) and two different target variables (i.e., trip distance and usage time) used for the evaluation of synthetic data utility for the training of supervised ML models. Table 1 presents an overview of the data set and a description of the variables. Finally, we divide the data into training and test sets using an 80/20 split.

Table 1
Overview of explanatory and target variables.

Category	Variable	Description	Values	Mean SD	Min Max
Time	Time of day	Time of day the trip starts (time intervals, e.g., early morning)	6	–	–
	Day	Day of the week the trip starts (e.g., Monday)	7	–	–
	Non-working day	The trip starts on a weekend or public holiday (i.e., 1 or 0)	2	–	–
	Month	Month the trip starts (e.g., July)	12	–	–
Location	District	District the trip is started from (e.g., downtown)	21	–	–
Car	Vehicle ID	Unique ID of every vehicle in the fleet (e.g., 101)	13	–	–
User	Age	Age of the user at the time of the trip (in years)	–	34.35 11.40	17 86
	Gender	Gender of the user (i.e., female, male, or not specified)	3	–	–
	Contract type	Type of contract of the user (e.g., private, business, or internal)	13	–	–
	Contract duration	Duration of the user's contract until the start of the trip (in months)	–	19.86 16.70	0 79
	Business	Indicator for business clients (i.e., 1 or 0)	2	–	–
	Student	Indicator of the student status of the user (i.e., 1 or 0)	2	–	–
	Public transport	Indicator for a subscription to local public transport (i.e., 1 or 0)	2	–	–
Prior Usage	Prior trips	Aggregated number of trips by the user until the start of the trip (absolute number)	–	18.34 25.37	0 154
	Prior usage time	Mean prior usage time per trip of the user (in minutes)	–	111.36 161.83	0 4100.5
	Prior distance traveled	Mean prior distance traveled per trip of the user (in km)	–	15.07 18.46	0 401
	Prior drop-off distance	Mean prior distance between start and end points per trip of the user (in m)	–	998.95 1127.49	0 15,487
Weather	Weather ID	Main weather condition based on a group of weather parameters at the start of the trip (e.g., cloudy)	5	–	–
	Temperature	Perceived temperature (windchill factor) at the start of the trip (in degrees Celsius)	–	9.59 9.28	−18.28 30.9
	Rain	Rain volume in the last hour before the start of the trip (in mm)	–	0.11 0.38	0 6.67
Trip (target variables)	Trip distance	Distance traveled per trip of the user (in km)	–	16.26 20.67	1 199
	Usage time	Usage time per trip of the user (in minutes)	–	115.18 139.76	2 1065

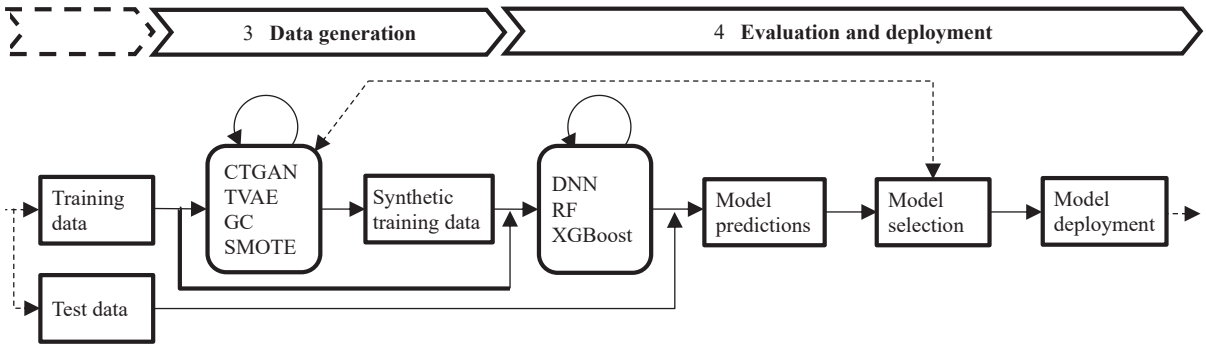


Fig. 4. Conceptual overview of the research phases 3 and 4.

Third, we proceed with synthetic *data generation*. Fig. 4 provides a conceptual overview of the steps in this and the subsequent research phase. We draw on conditional tabular GAN (CTGAN) and tabular VAE (TVAE) (Lei et al., 2019) as state-of-the-art generative ML algorithms to create synthetic data sets based on the training data. These models are specifically designed for the creation of synthetic tabular data and have lately been successfully employed in various domains (He and Zhou, 2022; Inan et al., 2023). Compared to standard GANs, CTGAN incorporates mode-specific normalization and conditional training–by–sampling to be able to deal with tabular data. Mode-specific normalization enables the modeling of multi-modal distributions in numerical columns. It uses a variational GMM (Bishop, 2006) to determine the number of modes per column and normalize its values according to the estimated distributions. The encoded values are then used during model training before being transformed back to the original scale for the synthetic data generated. Conditional training–by–sampling addresses imbalanced category-level frequencies by encoding categorical variables into condition vectors, sampling them according to the log frequency of the categories to incorporate rare categorical levels, and using them as generator inputs and as a filtering condition for sampling from the real data distribution (Lei et al., 2019). To improve learning stability and data quality, CTGAN leverages recent advances in GAN training from PacGAN (Lin et al., 2020) and Wasserstein GAN with gradient penalty (Gulrajani et al., 2017). The TVAE proposed by Lei et al. (2019) is an adaption of the standard VAE to fit tabular data. It learns the latent space distribution of the real training data before attempting to replicate the data while minimizing evidence lower-bound (ELBO) loss. For both models, we add sampling conditions to capture the business logic of the real data. We tune the models' hyperparameters using random grid search (Bergstra and Bengio, 2012) and 5-fold cross-validation (Stone, 1974) on the training data. We complement the generative ML models with a Gaussian Copula (GC) model (Li, 1999; Patki et al., 2016) and synthetic minority oversampling technique-nominal continuous (SMOTE–NC) (Chawla et al., 2002) as benchmark algorithms. For all models, we generate one data set (1) to replace the real training data, one data set (2) to rebalance selected explanatory variables in the real training data (i.e., 'contract type' and 'district'), and several data sets of different sizes (3) to augment the real training data with synthetic samples (i.e., adding 0.5, 1, 2, 4, and 8 times the original amount of data). We choose this wide range of augmentation factors as the optimal size of training data for ML models is difficult to determine in advance and often depends on an ensemble of factors such as model complexity, data diversity, and task specifics.

Fourth, we conclude with the *evaluation and deployment* phase where we aim to assess the quality of the synthetic data, select the best generative model for the present use case, and derive implications for the deployment of the model in the real-world setting. We start with basic quality checks on the synthetic data sets (e.g., avoidance of real data duplicates, adherence to the data boundaries, and coverage of all categories) before evaluating the fidelity of the synthetic data sets (i.e., the statistical similarity to the real data). To this end, we use descriptive statistics and graphical representations to evaluate the similarity of the marginal distributions. We draw on the complement to the two-sample Kolmogorov–Smirnov (KS) statistic for numerical variables (Massey, 1951) and on the complement to the Total Variation Distance (TVD) for categorical variables as distance metrics. We further assess pairwise Pearson correlations of the data sets' numerical variables. Next, we evaluate the utility of the synthetic data sets (i.e., the suitability for training ML models) against the real training data baseline using three supervised ML algorithms (i.e., DNN, RF, and extreme gradient boosting (XGBoost)). For this purpose, we employ two evaluation protocols that correspond to the prediction of two different target variables (i.e., trip distance and usage time). In each case, we evaluate the performance of the prediction models for (1) replacing the real training data with synthetic data, (2) rebalancing the real training data with synthetic data, and (3) augmenting the real training data with synthetic data. For validation, we also investigate (2) and (3) using duplicates of the real training data for rebalancing and augmenting the real training data. We evaluate the model performance on the test data for the different training data sets drawing on mean average error (MAE): $MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{Y}_i - Y_i|$, where Y_i are the observed values, $\widehat{Y}_i$ are the predicted values, and n is the number of observations. We additionally report the root mean squared error (RMSE). Based on the results, we finally select the generative model that corresponds to the synthetic training data yielding the best prediction accuracy (i.e., the highest ML efficacy and utility for the given prediction tasks) and derive implications for its deployment.

4. Results

In this section, we present the results and key insights of our study with a particular focus on the evaluation phase of our research approach. We first evaluate the fidelity of the synthetic data sets and thus determine to what extent they accurately represent the structural characteristics of the real data. Table 2 provides an overview of the data sets' descriptive statistics while Fig. 5, Fig. 6, and Fig. 7 support the evaluation with graphical representations. Thereafter, we assess the utility of the synthetic data sets by analyzing if and how they can support the prediction tasks of the present use case, where the operator aims to obtain more reliable predictions of the trip distance and the usage time as a basis for its operational and strategic decisions for the expansion of its free-floating EV service. Table 3 presents the prediction performances for the two evaluation protocols as the basis for model selection.

The fidelity of the synthetic data sets created by the four generative models (i.e., CTGAN, TVAE, GC, and SMOTE-NC) is first validated based on their statistical characteristics as indicators for similarity to the real training data. To this end, we compare the means, medians, standard deviations (SD), minima (min), and maxima (max) for the target variables as well as the overall compatibility of distribution functions and pairwise correlations of the real and synthetic data as presented in Table 2. In addition, we performed basic quality checks on the synthetic data sets to ensure that essential requirements were met (e.g., avoidance of real data duplicates, adherence to the data boundaries, and coverage of all categories). Overall, the statistical metrics indicate that all synthetic data sets are structurally similar to the original training data set. However, some peculiarities require closer analysis.

Regarding the means and medians of the two target variables, the data generated by CTGAN, TVAE, and SMOTE-NC closely resemble the original data. In contrast, the mean and median of the GC data are significantly higher than expected. In terms of the SDs of the two target variables, the GC data also exhibit the greatest differences from the original data. Analyzing the min and max values of the data sets reveals that CTGAN and TVAE nearly cover the full range of the original data. However, the benchmark models' max values do reach those of the real data. For the 'trip distance', the CTGAN data provide the closest resemblance to the real data, with their metrics being almost identical. Fig. 5 supplements the descriptive statistics with a graphical representation of the marginal distributions of the 'trip distance' for all synthetic data sets to allow a visual assessment. It confirms both the shift of the mean of the GC data and the restricted range of the data generated by both benchmark models. When it comes to the 'usage time', the SMOTE-NC data demonstrate the greatest statistical similarity to the real data, closely followed by the CTGAN and TVAE data. In addition to the inspection of the target variables, Fig. 6 exemplarily visualizes the marginal distributions of selected explanatory variables for the CTGAN data and the real training data.

We further evaluate the similarity of the distributions of numerical variables by computing the maximum differences between the cumulative distribution functions (CDF) of the real data and the synthetic data sets. Table 2 shows the average KS complement across all numerical variables and, analogously, the average TVD complement across all categorical variables. It can be seen that in this case, the CTGAN and GC data do not keep up with the other two data sets. Table 4 in the Appendix gives a detailed overview of the KS and TVD complements per variable. We observe that variable complexity (e.g., the number of categories in categorical variables) strongly affects the similarity of the marginal distributions to the original data. Accordingly, the marginal distributions of categorical variables such as 'Contract type' and 'Vehicle ID' in most synthetic data sets show a higher deviation from those of the real data.

As a final check of the structural similarity, we examine the differences in pairwise correlations between the real and synthetic data sets. We find similar pairwise correlations of the numerical variables to the real data for all synthetic data sets. This suggests that the synthetic data captures important aspects of the correlation structure observed in the real data. Fig. 7 exemplarily illustrates the pairwise correlations of numerical variables of the real and CTGAN data in the form of parallel heat maps. From the evaluation of data fidelity, we conclude that the synthetic data sets of all analyzed models adequately captured the distribution of the real training data. However, this only seems to be true to a limited extent for the GC data. We find that deviations in data quality are indicated by both descriptive statistics and graphical representations.

Thereafter we evaluate the utility of the synthetic data. Thus, we aim to test how well the generated synthetic data fit the purpose of enhancing the database for the training of prediction models and how they affect their performance. Table 3 presents the results for the

Table 2
Statistical metrics of the real and synthetic data sets.

	Metric	Real data	CTGAN	TVAE	GC	SMOTE-NC
Trip distance	Mean	16.04	15.93	14.39	18.26	14.47
	Median	10	10	9	16	10
	SD	19.73	19.88	17.90	15.85	15.88
	Min	1	1	1	1	1
	Max	197	196	192	87	173
Usage time	Mean	115.32	86.84	93.09	131.82	106.30
	Median	66	49	56	118	64
	SD	140.45	122.30	117.30	114.04	125.20
	Min	2	2	2	2	2
	Max	1065	940	1065	606	919
	KS complement (numerical variables)		0.8972	0.9234	0.8167	0.9410
	TVD complement (categorical variables)		0.8825	0.9073	0.9046	0.9203

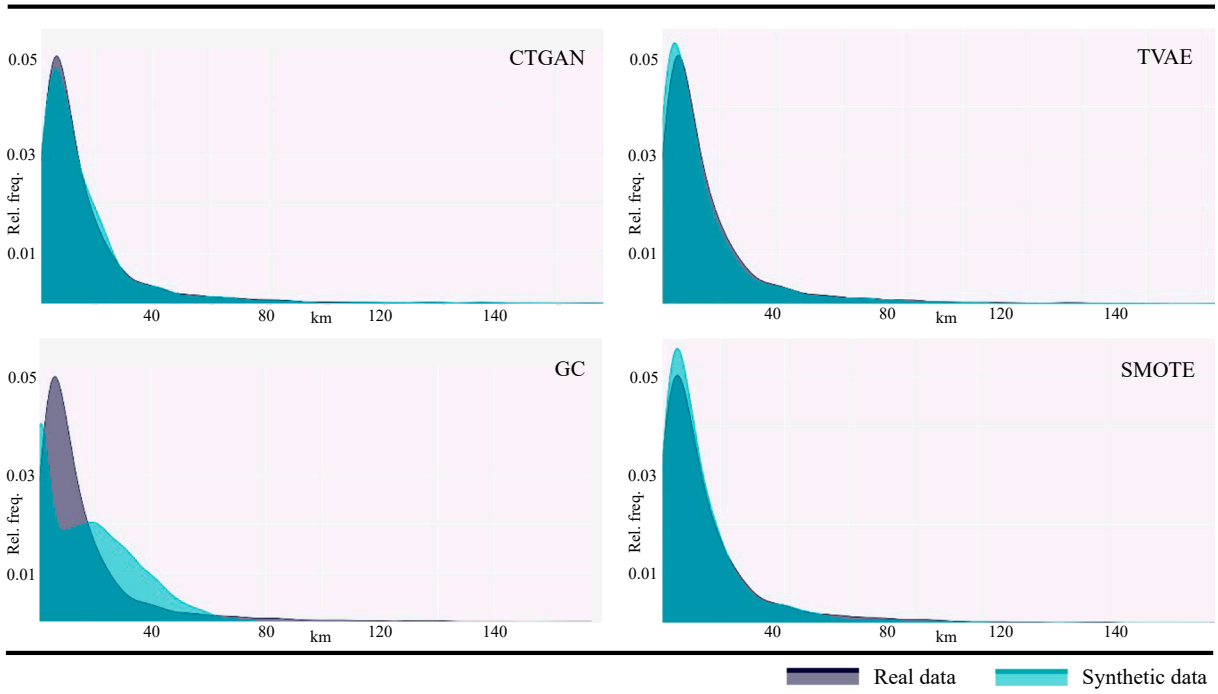


Fig. 5. Marginal distributions of the target variable 'trip distance'.

prediction accuracy across all generative models and prediction models for the two evaluation protocols (i.e., the target variables 'trip distance' and 'usage time') next to those of the real data.

In this evaluation phase, we first examine the results of *replacing* the real training data with synthetic data sets of the same size (i.e., 4756 trips). Regarding the prediction of the 'trip distance', training the prediction models with the CTGAN data yields the best average results, closely followed by the TVAE data which performs better than the SMOTE-NC and GC data. Except for the latter, the average prediction results across all prediction models using the synthetic training data are within an MAE range of 0.3457 or 3.01 % compared to the average results using the real training data. For the CTGAN and TVAE data, the results per single prediction model do not exceed an MAE difference of 0.5151 or 4.35 % to the results achieved when training the same models with real data. As for the benchmark models, the results per single prediction model are within an MAE range of 0.7089 or 6.22 % for the SMOTE–NC data and 2.2032 or 19.56 % for the GC data when compared to the results with real data.

For the prediction of the 'usage time', the best average results are achieved with the TVAE data just ahead of the CTGAN data. For these generative ML models, the average results across all prediction models using the synthetic training data are within an MAE range of 1.3123 or 1.64 % compared to the average results using the real training data. In contrast, the average MAE differences of the benchmark models are 2.2374 or 2.80 % for the SMOTE-NC data and 11.5546 or 14.45 % for the GC data. When comparing the results of the real and synthetic data per prediction model, the MAE difference does not exceed 2.8241 or 3.57 % for both generative ML model data sets, 3.1283 or 3.97 % for the SMOTE–NC data, and 14.1015 or 17.83 % for the GC data. Overall, we observe similar prediction accuracies when replacing the real training data with synthetic data sets of the same size. The CTGAN and TVAE data lead to the most promising results while the GC data lag behind the other data sets.

In the next step, we analyze the outcomes of *rebalancing* selected explanatory variables in the real training data (i.e., 'contract type' and 'district') by purposefully adding synthetic data containing underrepresented values of these categorical variables (i.e., 1564 synthetic trips added to 4756 real trips). For the prediction of the 'trip distance', rebalancing the original data using synthetic data created by the CTGAN and TVAE improves the training and accuracy of all prediction models investigated (i.e., six times out of six). This translates into an average improvement of 2.57 % for the CTGAN data and 0.67 % for the TVAE data across all prediction models compared to only using the real training data. In contrast, employing the benchmark model data for the same task yields a smaller improvement of 0.1740 % for the SMOTE-NC data and a decline of 0.10 % using the GC data. Overall, rebalancing the training data with synthetic samples always leads to better prediction results than rebalancing with real data duplicates, which does not improve performance compared to using only the real training data.

For the prediction of the 'usage time', the results are even more distinct. Rebalancing the original data with CTGAN and TVAE data improves the prediction accuracy on average by 1.18 % and 0.44 %. This corresponds to an improvement over using the original training data alone in six of six cases. This is true in only one of six cases for the data generated by the benchmark models where the rebalancing results in an average performance decline of 0.75 % considering all prediction models. In this connection, rebalancing the training data with synthetic samples only leads to better prediction results than rebalancing with real data duplicates when using CTGAN and TVAE data. Overall, we find that rebalancing certain variables in the original training data by supplementing conditional

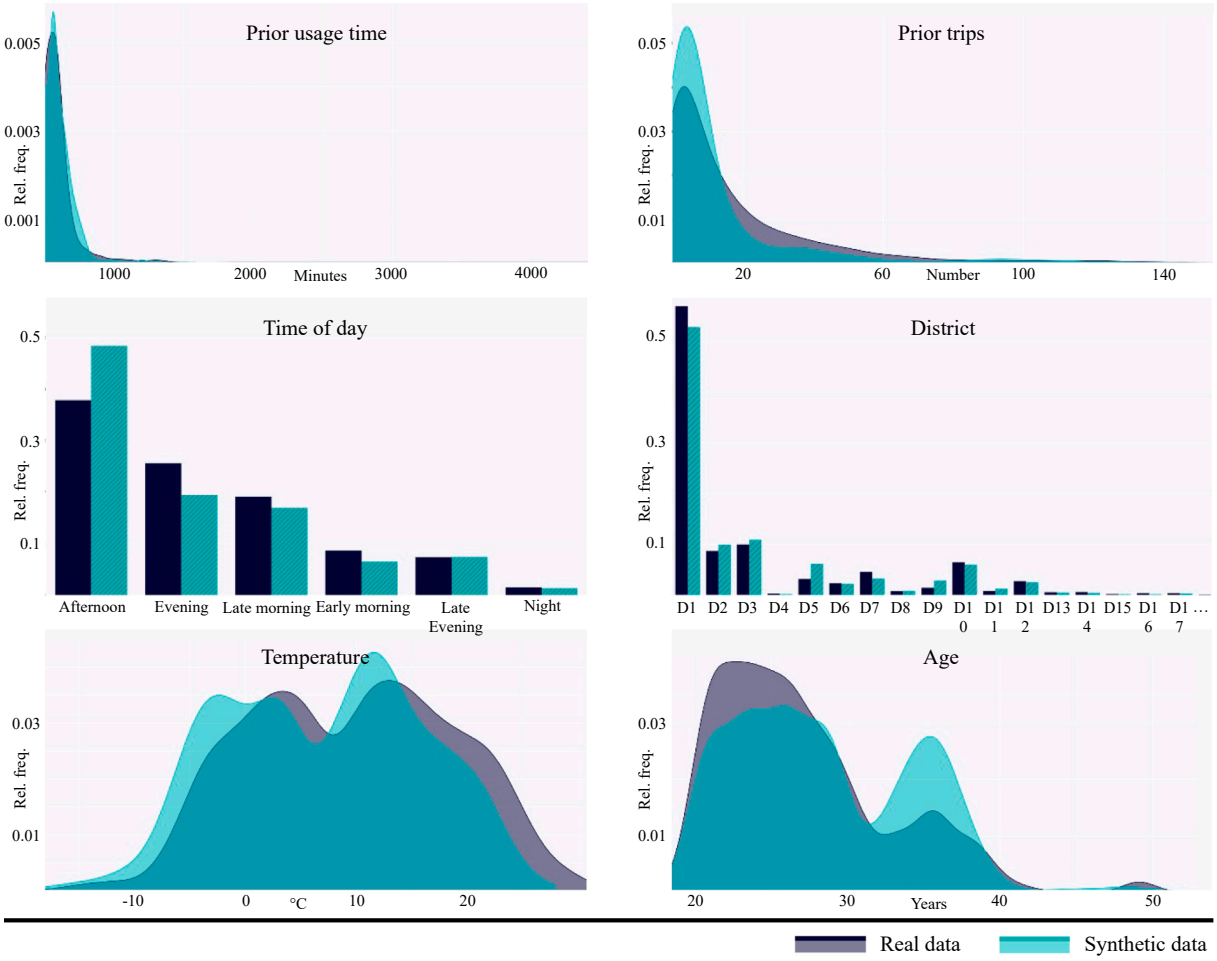


Fig. 6. Marginal distributions of selected explanatory variables (CTGAN).

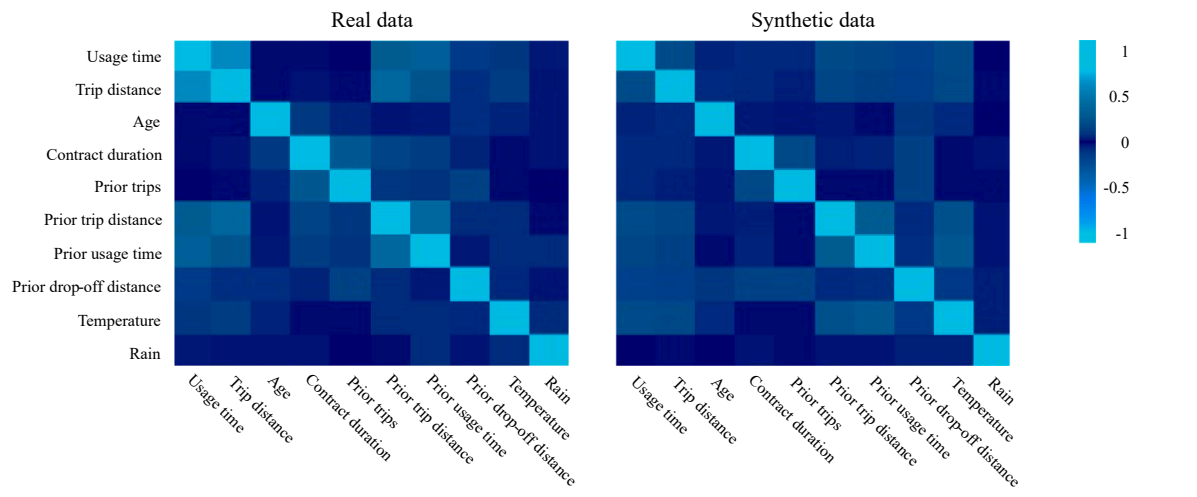


Fig. 7. Heat maps of the pairwise correlations of numerical variables (CTGAN).

Table 3
Prediction accuracy for the synthetic training data sets and the real data (MAE and RMSE).

Target variable	Evaluation mode	CTGAN data			TVAE data			GC data			SMOTE-NC data			Real data		
		NN	**RF**	**XGBoost**	**NN**	**RF**	**XGBoost**	**NN**	**RF**	**XGBoost**	**NN**	**RF**	**XGBoost**	**NN**	**RF**	**XGBoost**
Trip distance	replacing	11.3196	11.5520	11.4922	11.8184	11.3487	11.5896	12.3650	13.4651	13.4392	11.8364	11.5884	12.0987	**11.8347**	**11.2619**	11.3898
		23.2087	22.6227	22.5921	22.0950	21.8802	21.8726	22.5965	22.8592	22.4730	21.8198	22.2567	22.2128	20.9877	21.2480	21.0495
	rebalancing	11.3523	11.0855	11.1611	11.7991	11.2125	**11.2455**	11.8883	**11.2598**	11.3729	11.6874	11.2754	11.4637	12.7694	11.3134	**11.3543**
		21.2471	21.4888	21.2945	21.1358	21.3523	21.0381	21.0460	21.3563	21.2242	21.3519	21.4206	21.1773	21.3496	21.3229	21.3265
	augmenting	**11.0701**	**10.9784**	**10.8630**	**11.2406**	**11.1070**	11.4500	**11.6539**	11.3464	**11.1518**	**11.6143**	**11.2061**	**11.1693**	12.5649	11.2968	11.5540
		22.3523	21.4341	22.7119	21.5538	21.2524	21.2246	21.6236	21.4064	21.3925	21.4031	21.2714	21.2983	22.5950	21.3258	21.1755
Usage time	replacing	81.2940	81.9108	80.6387	81.4213	77.6522	**77.4637**	87.6569	93.1882	93.7253	84.2916	80.5134	81.8139	**82.1344**	79.0867	**78.6856**
		128.0502	126.5895	127.8963	129.2339	124.1547	124.8255	128.5637	129.1983	129.4977	131.1395	123.2877	124.5412	122.8217	119.5922	119.2367
	rebalancing	80.0625	78.8520	78.1554	81.7888	78.5501	78.5069	84.6442	**79.2452**	**78.6709**	82.6343	79.2314	**79.0084**	82.1817	**78.5650**	78.9757
		119.9833	120.1922	119.5669	124.1220	118.5909	118.3571	122.7949	119.6962	119.2687	124.0749	119.7012	119.6470	125.0647	118.2102	119.6662
	augmenting	**79.8870**	**77.9111**	**76.4161**	**80.7069**	**76.1038**	77.9207	**78.7016**	79.4261	81.0885	**84.0167**	**77.8619**	79.4226	87.4603	78.7227	80.3747
		121.2118	118.2622	118.7290	123.9520	118.5167	118.8860	123.7068	119.0203	120.0561	129.6483	119.4801	120.4374	130.4195	119.7439	119.4032

synthetic data samples created by the generative ML models improves the training and accuracy of prediction models. For the benchmark model data, this applies to only one third of the constellations investigated.

Finally, we analyze the impact of *augmenting* the real training data with synthetic data to increase the number of training samples (i.e., 2378, 4756, 9512, 19024, or 38,048 synthetic trips added to 4756 real trips). Table 5 in the Appendix presents the complete results for all augmentation quantities while Table 3 reports the best accuracy achieved per synthetic data set and prediction model. Starting with the prediction of the 'trip distance', we find that augmenting the original data set with synthetic data improves the training and accuracy of the prediction models in 10 out of 12 cases evaluated. This corresponds to an average improvement of the MAE of 0.5250 or 4.57 % for the CTGAN data across all prediction models compared to the training with the original data. We further find an average improvement of the MAE of 0.2297 or 2.00 % for the TVAE data, 0.1656 or 1.44 % for the SMOTE-NC data, and 0.1115 or 0. 97 % for the GC data.

Considering the prediction of the 'usage time', the data augmentation yields accuracy improvements in all of the prediction tasks supported with generative ML data and in 8 of 12 cases overall. The average MAE improvement across all prediction models compared to the training with the original data is 1.8975 or 2,37 % for the CTGAN data, 1.7251 or 2.16 % for the TVAE data, 0.6390 or 0.80 % for the SMOTE-NC data, and 0.2302 or 0.29 % for the GC data. Regarding the optimal augmentation quantity and ratio respectively, we observe that, in our setting, augmenting the training data by adding 0.5 to 2 times the original amount of training data (i.e., 2378 to 9512 synthetic trips added to 4756 real trips) mostly yields the greatest performance improvement for all prediction models. We achieve the highest overall improvements with an augmentation factor of 2. As for the benchmark model data, the lower quality of synthetic data may imply a lower optimal augmentation quantity. Augmenting the real training data through duplication does not improve performance. From the results of both evaluation protocols, it is evident that increasing the amount of training data alone does not improve model training and prediction performance. Instead, the quality of synthetic data as well as the ratio of synthetic to real data are critical to maximizing data utility for ML prediction tasks.

We conclude that the synthetic data sets created by the generative ML models offer the greatest data utility for the prediction tasks at hand. In this respect, the CTGAN data have the edge over the TVAE data while both models outperform the benchmark models regarding average performance across all prediction models as well as best individual prediction results. Overall, comparing the best prediction results of the real training data baseline with those using the augmented training data across all models shows an improvement in prediction accuracy of 3.54 % for the 'trip distance' as the target variable and 3.28 % for the 'usage time'. Opposed to the results for the same prediction models, these improvements rise to 4.63 % and 3.77 %. These results show that the training of prediction models can be notably improved by augmenting real training data with synthetic data to gain more accurate insights on upcoming trips. These improvements could not be achieved by duplicating the available real data.

Although not the focus of this study, we also note that XGBoost yields the overall best prediction results and that, as expected for the present tabular data, the tree-based models outperform the NN. Considering the previous insights on the fidelity of the synthetic data sets, we find that basic descriptive statistics can be useful indicators for estimating the quality of synthetic data. As such, the statistical metrics previously indicated that the CTGAN data closely represent the 'trip distance' which translated to exceptionally high data utility for the prediction of this target variable. At the same time, the descriptive statistics also suggested a lower quality of the GC data which was confirmed by the utility evaluation. However, the KS and TVD complements as metrics for the overall compatibility of the distribution functions are not found to be accurate indicators of data utility for the present prediction tasks.

In the present case, the operator aims to specifically make use of the finding for planning the charging cycles of its EVs (i.e., provide users with real-time suggestions of charging stations and times based on the current battery level of the EV and predictions of the trip distance) and for optimizing its booking system (i.e., provide users with vehicle availability forecasts despite the open-ended nature of EV bookings). While the integration of models into the existing infrastructure and systems of the operator (i.e., systems engineering) is out of the scope of this work, we share implications for model deployment for the present case based on the above findings. For the initial deployment, we propose to augment the available trip data with synthetic data generated by CTGAN to train a tree–based prediction model such as XGBoost. After the initial deployment, it is necessary to be able to regularly update the models to make sure they operate on the most recent database and include all available real data (e.g., scheduled regular retraining). In this connection, we suggest employing an expanding window that incorporates new training data as it becomes available. The ideal amount of synthetic training data augmenting the real data should then be evaluated routinely. In addition, both generative and prediction models need to be updated regularly to incorporate potential changes in the joint distributions (i.e., concept drift). In the course of this, standard ML monitoring objectives such as data integrity checks, anomaly detection, and (re-)evaluation of performance metrics should be implemented.

5. Discussion

The results of our study present novel insights into the use of generative ML for the creation of synthetic mobility data. The findings show that augmenting real training data with synthetic samples created by generative ML models improves predictive modeling of upcoming trips when data are scarce. This study helps understand the methods and approaches for effectively generating and leveraging synthetic tabular transaction data of carsharing trips for more accurate predictions of user behavior. Thus, our work has implications that advance current research on decision support and the introduction of new services in carsharing and provide novel insights for decision–makers in practice (e.g., carsharing operators, municipal planners, and policy makers).

First, this research presents a new application-oriented perspective on the use of generative ML in transportation research and stimulates further exploration of leveraging synthetic tabular mobility data. Recent research highlights the multi-faceted potential of generative ML for addressing transportation problems (Wong and Farooq, 2020). While most studies primarily focus on tailored methodological solutions for simulation tasks such as trajectory modeling (Zhang et al., 2019; Choi et al., 2021) and population synthesis (Garrido

et al., 2020; Kim et al., 2022), research has also started to analyze the use of synthetic training samples for prediction tasks (Boquet et al., 2020; Li et al., 2020). Our work contributes to this body of knowledge by being the first to investigate generative ML algorithms able to create synthetic tabular data (i.e., CTGAN and TVAE) to overcome the barriers faced during the introduction and expansion of innovative mobility options. The results of our study demonstrate that synthetic data can help to quickly reach the necessary database to obtain valuable predictions of user behavior. This adds a technological and data dimension to research on business aspects of carsharing in less-densely populated areas (Wappelhorst et al., 2014; Rotaris and Danielis, 2018; Baumgarte et al., 2022) and in emerging economies (Luna et al., 2020) as well as on new carsharing business models and services (Lagadic et al., 2019; Zhang et al., 2020). In addition, our study presents a new application–oriented perspective on synthetic data creation using generative ML models by going through an iterative ML workflow including model evaluation, selection, and implementation based on a real–world case of an emerging carsharing program. The results highlight the high potential of state–of–the-art generative ML algorithms like CTGAN and TVAE for creating high-quality synthetic tabular data. Thus, we hope to present a blueprint to fellow scholars for investigating the use of synthetic mobility data in a broader range of domains and advance previous research building on statistical or simulation approaches to generate synthetic data (Etxandi-Santolaya et al., 2023). We are confident that the insights of this study can be valuable to future research on other innovative mobility services (e.g., bike sharing, e-scooter sharing, or ride–hailing).

Second, our results contribute to research on carsharing decision support by investigating approaches to effectively generate and leverage synthetic trip data to mitigate the challenge of limited data access. In the increasingly competitive market of innovative mobility, insights into user behavior and reliable predictions of trip characteristics through ML models have become indispensable for carsharing operators as well as policy makers (Cheng et al., 2021; Baumgarte et al., 2022). Previous literature is primarily concerned with developing methodological solutions to the prevalent decision support problems in carsharing (Lei et al., 2020; Wang et al., 2021; Prinz et al., 2022). However, restricted access to enough high-quality real-world data is still one of the main limitations to the application and broad evaluation of many innovative models (Brendel et al., 2017; Lu et al., 2022). In this connection, the results of our study show that synthetic carsharing transaction data generated by GANs and VAEs improve ML training and model accuracy when the amount or quality of data is inadequate. The presented insights can advance prescriptive carsharing research that leverages predictions of user behavior, inter alia, for the optimization of fleet management, pricing schemes, and charging policies (Perboli et al., 2018; Xu and Meng, 2019). In addition, our research particularly connects to recent studies concerned with the practical issues faced by small and emerging carsharing programs that aim to break new ground and challenge established enterprises (e.g., by opening up small urban areas, offering all-EV fleets, or launching municipal programs), but lack a long-term data record (Rotaris and Danielis, 2018; Lagadic et al., 2019; Illgen and Höck, 2020). Our study assists fellow scholars with creating and using synthetic mobility data for developing solutions to real–world carsharing decision support problems. In this connection, it also specifically contributes to research on the introduction and expansion of EV carsharing services (e.g., the planning of charging cycles and infrastructure based on anticipated user behavior) (Cocca et al., 2019; Luna et al., 2020).

Third, our work supports carsharing decision-makers in leveraging predictions of trip characteristics by providing a solution to limited data access for small and emerging operators. The results of our study show that the training of prediction models can be improved by up to 4.63 % by augmenting real training data with synthetic samples when data are scarce. To put these results into practice, this research presents a feasible way for carsharing operators to evaluate and deploy generative ML models for qualitatively and quantitatively enhancing their database. This helps operators to obtain more accurate prior insights on trip characteristics during the introduction or expansion of their services where data access is a common challenge (Brendel et al., 2017; Lagadic et al., 2019). Such predictions are a valuable basis for operational and strategic decisions that enhance the overall customer experience or optimize business operations. In this regard, the results can contribute to bringing the technological capabilities of small and emerging operators and large carsharing companies closer together. More accurate predictions of trip distances and usage times during booking can be used, for instance, to provide customized service offerings in the booking process (e.g., individualized coupons or vehicle suggestions like switching to an EV for short trips), to incentivize EV charging activities performed by the users, or to develop dynamic pricing models that align with the predicted trip distance and usage time (Brendel et al., 2018; Cocca et al., 2019; Hu et al., 2021). Furthermore, operators can leverage accurate predictions of trip characteristics to support the planning of vehicle maintenance and inspections as well as to assist fleet rotation and replacement decisions which optimizes the utilization of vehicles and reduces the risk of breakdowns. In addition, high-quality synthetic data can contribute to a closer integration of carsharing and the public sector (e.g., municipalities) by facilitating data sharing and informed policy decisions. Robust public–private partnerships in turn foster policies that help make carsharing as a sustainable mobility option more attractive (e.g., expansion of the public charging infrastructure or designation of free parking zones) (Lagadic et al., 2019; Vanheusden et al., 2022). Finally, the results of this study cater to municipal carsharing programs which are particularly subject to strict data protection regulations that hinder data acquisition and transfer.

6. Conclusion

This study explored the creation and use of synthetic data to support carsharing decision-making by overcoming the barrier of limited data access during the introduction and expansion of new services, enabling more accurate predictions of trip characteristics. It investigated the evaluation, selection, and implementation of state-of-the-art generative ML models (i.e., GANs and VAEs) to create synthetic tabular transaction data of carsharing trips. To this end, it examined the case of an emerging municipal carsharing program that is expanding its services to free–floating EVs but lacks the database to obtain reliable trip predictions.

The results of this study show that augmenting real training data with synthetic samples improves the performance of prediction models by up to 4.63 % when predicting the usage time and distance of upcoming trips. In practice, these improvements are significant, as they mean that operators can, for instance, more accurately predict upcoming trip distances by more than 500 m. The findings of the

analysis further reveal that the quality of synthetic data (i.e., using generative ML models such as CTGAN and TVAE compared to statistical benchmark models) as well as the ratio of synthetic to real data (i.e., using reasonable augmentation factors of 0.5 to 2) are critical to maximizing data utility for ML prediction tasks. The results present novel insights on the use of generative ML for the creation of synthetic mobility data and help understand the methods and approaches for leveraging synthetic tabular transaction data of carsharing trips for more accurate predictions of user behavior. Carsharing operators can draw on our study to enhance their available database when launching or expanding their services to achieve more accurate predictions of upcoming trips and align their service offers with anticipated user behavior.

The present study is constraint by the following limitations that may point fellow scholars in the direction of further beneficial research. First, it should be noted that synthetic data and associated results should be treated with caution. Depending on the real-world database, model choice, and technical fine-tuning (e.g., model architecture, training procedure, and loss functions) the synthetic data may not fully capture the complexity and diversity of actual data, reproduce biases, and potentially compromise model generalizability to real-world scenarios. Further, the results and implications of this work are developed based on the case of one emerging carsharing program. Thus, the findings need to be validated for other use cases and environments (e.g., station-based services or completely new carsharing programs) to include potential contextual constraints. Nevertheless, we consider the application-oriented approach and real-world data to be a strength of this study and we are confident that the presented research design can serve as a blueprint to address this limitation in future research. In addition, future studies should build on our results and investigate the actual impact of the achieved improvements in prediction accuracy on specific decision-making areas and the strategic planning of carsharing operators. Finally, fellow scholars are encouraged to expand the scope of the present research to other generative ML models as well as to investigate the privacy benefits of synthetic data in carsharing and transportation research.

CRediT authorship contribution statement

Tobias Albrecht: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing, Software. **Robert Keller:** Resources, Supervision. **Dominik Rebholz:** Data curation. **Maximilian Röglinger:** Resources, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Table 4
KS and TVD complements per variable.

	Variable	CTGAN	TVAE	GC	SMOTE-NC
Numerical variables	Usage time	0.8415	0.9205	0.791	0.9668
	Trip distance	0.9746	0.9308	0.7719	0.9704
	Age	0.8618	0.9155	0.9464	0.9076
	Contract duration	0.951	0.882	0.9022	0.9367
	Prior trips	0.8358	0.9212	0.8057	0.9357
	Prior trip distance	0.8898	0.909	0.7937	0.9624
	Prior usage time	0.9119	0.9302	0.7803	0.9695
	Prior drop-off distance	0.9087	0.9048	0.8621	0.9834
	Temperature	0.8921	0.9426	0.9546	0.9319
	Rain	0.9045	0.9769	0.5591	0.8452
Mean		**0.8971**	**0.9234**	**0.8167**	**0.9401**
Categorical variables	Time of day	0.894	0.8158	0.8778	0.886
	Day	0.8516	0.8955	0.8877	0.9527
	Non-working day	0.9304	0.9729	0.9533	1.000*
	Month	0.8106	0.9035	0.8909	0.9638
	Vehicle ID	0.8293	0.8942	0.8884	0.9765
	District	0.9296	0.7075	0.9008	0.8139
	Contract type	0.7847	0.9058	0.8671	0.8806
	Business	0.9153	0.975	0.939	0.9447
	Public transport	0.9817	0.9933	0.9527	0.9487
	Student	0.9497	0.9809	0.9048	0.9073
	Gender	0.8793	0.9275	0.9026	0.8862
	Weather	0.8341	0.9159	0.89	0.8833
Mean		**0.8825**	**0.9073**	**0.9046**	**0.9131**

*Variable used to define the sampling strategy.

Table 5
Prediction accuracy per augmentation factor for the real and synthetic training data sets (MAE and RMSE).

Target variable	Augment. factor	CTGAN data			TVAE data			GC data			SMOTE-NC data			Real data		
		NN	RF	XGBoost	NN	RF	XGBoost	NN	RF	XGBoost	NN	RF	XGBoost	NN	RF	XGBoost
Trip distance	0.5	11.5518	11.5190	11.2139	**11.2406**	11.2108	11.5628	**11.6539**	**11.3464**	**11.1518**	11.6383	**11.2061**	**11.1693**	**12.5649**	11.3533	11.5847
		21.6734	21.5736	21.3607	21.5538	21.3332	21.2258	21.6236	21.4064	21.3925	20.9795	21.2714	21.2983	22.5950	21.2792	21.2970
	1	12.3415	11.0811	11.1880	11.6267	11.3852	**11.4500**	11.6832	11.4623	11.7388	11.8701	11.4504	11.9608	13.5081	**11.2968**	**11.5540**
		21.3206	21.5401	21.4465	22.5738	21.5375	21.2246	21.5824	21.5770	21.3079	22.4179	21.2890	22.2181	23.5036	21.3258	21.1755
	2	**11.0701**	**10.9784**	**10.8630**	11.2944	**11.1070**	11.6146	12.2842	11.6670	11.9056	**11.6143**	11.3761	11.4569	13.9806	11.3331	11.5839
		22.3523	21.4341	22.7119	21.5708	21.2524	21.5624	21.6239	21.5522	21.5053	21.4031	21.4581	21.3131	23.6211	21.3885	21.1954
	4	11.2258	11.0166	11.0845	11.6463	11.1518	11.5070	12.4059	11.7770	12.0858	12.0885	11.3245	11.4488	15.1953	11.3385	11.5855
		21.7501	22.0809	21.7518	21.6216	21.4237	21.5522	21.4349	21.4929	21.5245	21.3162	21.4282	21.3659	24.7463	21.4240	21.1687
	8	11.1417	10.9786	11.0960	12.0344	11.2458	11.6128	12.6589	11.9042	12.3085	12.0435	11.4203	11.6557	15.2002	11.5697	11.6146
		22.6277	22.3583	21.8892	21.8159	21.2930	21.5330	22.0645	21.6740	21.6367	21.8965	21.5819	21.7079	24.7201	21.5069	21.2176
Usage time	0.5	82.2489	78.0861	**76.4161**	82.9419	76.6101	78.3930	86.5024	**79.4261**	**81.0885**	88.0047	79.5212	80.7611	**87.4603**	78.8165	80.7569
		122.2467	118.8730	118.7290	124.0266	117.6743	118.9320	123.7227	119.0203	120.0561	130.5965	120.8799	121.1762	130.4195	119.7548	120.1860
	1	85.3850	78.8599	79.9373	**80.7069**	**76.1038**	**77.9207**	**78.7016**	80.6413	82.2907	98.7119	78.9635	83.8295	94.8528	79.2820	80.7582
		120.8096	119.2592	119.5011	123.9520	118.5167	118.8860	123.7068	120.1383	118.9183	141.3106	119.6050	126.4877	138.2004	120.1050	119.8980
	2	**79.8870**	**77.9111**	79.0568	81.5373	76.9555	78.4206	89.9604	81.4714	81.6627	96.9908	78.8619	**79.4226**	98.0506	**78.7227**	**80.3747**
		121.2118	118.2622	118.1067	121.7598	120.0734	120.4933	124.2410	119.1872	119.6950	136.2353	119.2565	120.4374	143.4132	119.7440	119.4032
	4	81.9803	78.5853	78.8677	82.1284	76.9810	78.3428	93.3835	82.4530	81.3385	**84.0167**	**77.8619**	79.5495	106.7233	79.0091	80.5591
		119.0491	120.4995	118.4787	124.2911	120.4239	120.8641	126.0798	119.2925	119.2959	129.6483	119.4801	121.3269	149.3471	120.0924	119.7103
	8	80.4887	78.5772	79.5752	82.5054	76.2563	78.1685	90.9449	83.7901	86.4541	87.7515	78.7626	80.0561	108.0623	78.7574	80.9026
		124.4496	121.4946	121.4854	128.8035	118.5909	119.8460	124.3727	120.9385	120.8589	128.4495	121.4952	122.9563	155.0767	119.7945	120.2366

References

Abbasi, S., Ko, J., Kim, J., 2021. Carsharing station location and demand: Identification of associated factors through Heckman selection models. J. Clean. Prod. 279, 123846.

Albrecht, T., Rausch, T.M., Derra, N.D., 2021. Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivals' forecasting. J. Bus. Res. 123, 267–278.

Alencar, V.A., Pessamilio, L.R., Rooke, F., Bernardino, H.S., Borges Vieira, A., 2021a. Forecasting the carsharing service demand using uni and multivariable models. J Internet Serv Appl 12.

Alencar, V.A., Rooke, F., Cocca, M., Vassio, L., Almeida, J., Vieira, A.B., 2021b. Characterizing client usage patterns and service demand for car-sharing systems. Inf. Syst. 98, 101448.

Alqahtani, H., Kavakli-Thorne, M., Kumar, G., 2021. Applications of Generative Adversarial Networks (GANs): An Updated Review. Arch Computat Methods Eng 28, 525–552.

Amatuni, L., Ottelin, J., Steubing, B., Mogollón, J.M., 2020. Does car sharing reduce greenhouse gas emissions? Assessing the modal shift and lifetime shift rebound effects from a life cycle perspective. J. Clean. Prod. 266, 121869.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein Generative Adversarial Networks. International conference on machine learning. https://doi.org/10.48550/arXiv.1701.07875.

Baumgarte, F., Brandt, T., Keller, R., Röhrich, F., Schmidt, L., 2021. You'll never share alone: Analyzing carsharing user group behavior. Transp. Res. Part D: Transp. Environ. 93, 102754.

Baumgarte, F., Keller, R., Röhrich, F., Valett, L., Zinsbacher, D., 2022. Revealing influences on carsharing users' trip distance in small urban areas. Transp. Res. Part D: Transp. Environ. 105, 103252.

Bergstra, J., Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res. 13, 281–305.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer, New York, NY.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., 2013. A review of feature selection methods on synthetic data. Knowl Inf Syst 34, 483–519.

Boquet, G., Morell, A., Serrano, J., Vicario, J.L., 2020. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. Transportation Research Part C: Emerging Technologies 115, 102622.

Borji, A., 2022. Pros and cons of GAN evaluation measures: New developments. Comput. Vis. Image Underst. 215, 103329.

Boyacı, B., Zografos, K.G., 2019. Investigating the effect of temporal and spatial flexibility on the performance of one-way electric carsharing systems. Transp. Res. B Methodol. 129, 244–272.

Brahimi, N., Zhang, H., Dai, L., Zhang, J., 2022. Modelling on Car-Sharing Serial Prediction Based on Machine Learning and Deep Learning. Complexity 2022, 1–20.

Brendel, A.B., Rockenkamm, C., Kolbe, L.M., 2017. Generating Rental Data for Car Sharing Relocation Simulations on the Example of Station-Based One-Way Car Sharing. Proceedings of the 50th Hawaii International Conference on System Sciences, 1554–1563. https://aisel.aisnet.org/hicss-50/da/service_analytics/2/.

Brendel, A.B., Lichtenberg, S., Brauer, B., Nastjuk, I., Kolbe, L.M., 2018. Improving electric vehicle utilization in carsharing: A framework and simulation of an e-carsharing vehicle utilization management system. Transp. Res. Part D: Transp. Environ. 64, 230–245.

Brendel, A.B., Lichtenberg, S., Morana, S., Prinz, C., Hillmann, B.M., 2022. Designing a Crowd-Based Relocation System—The Case of Car-Sharing. Sustainability 14, 7090.

Brendel, A.B., Brennecke, J.T., Hillmann, B.M., Kolbe, L.M., 2023. The Design of a Decision Support System for Computation of Carsharing Pricing Areas and Its Influence on Vehicle Distribution. IEEE Trans. Eng. Manage. 70, 819–833.

Burghard, U., Dütschke, E., 2019. Who wants shared mobility? Lessons from early adopters and mainstream drivers on electric carsharing in Germany. Transp. Res. Part D: Transp. Environ. 71, 96–109.

car2go, 2019. car2go API Documentation. https://github.com/sharenowTech/openAPI Accessed 22 June 2023.

Carvajal-Patiño, D., Ramos-Pollán, R., 2022. Synthetic data generation with deep generative models to enhance predictive tasks in trading strategies. Res. Int. Bus. Financ. 62, 101747.

Chang, X., Wu, J., Correia, G.H.d.A., Sun, H., Feng, Z., 2022. A cooperative strategy for optimizing vehicle relocations and staff movements in cities where several carsharing companies operate simultaneously. Transport. Res. Part E: Logist. Transport. Rev. 161, 102711 https://doi.org/10.1016/j.tre.2022.102711.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Jair 16, 321–357.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets 14. http://arxiv.org/pdf/1606.03657v1.

Chen, T.D., Kockelman, K.M., 2016. Carsharing's life-cycle impacts on energy use and greenhouse gas emissions. Transp. Res. Part D: Transp. Environ. 47, 276–284.

Cheng, J., Chen, X., Ye, J., Shan, X., 2021. Flow-based unit is better: exploring factors affecting mid-term OD demand of station-based one-way electric carsharing. Transp. Res. Part D: Transp. Environ. 98, 102954.

Chicco, A., Diana, M., 2021. Air emissions impacts of modal diversion patterns induced by one-way car sharing: A case study from the city of Turin. Transp. Res. Part D: Transp. Environ. 91, 102685.

Choi, S., Kim, J., Yeo, H., 2021. TrajGAIL: Generating urban vehicle trajectories using generative adversarial imitation learning. Transportation Research Part C: Emerging Technologies 128, 103091.

Ciociola, A., Cocca, M., Giordano, D., Mellia, M., Morichetta, A., Putina, A., Salutari, F., 2017. UMAP: Urban mobility analysis platform to harvest car sharing data. In: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 1–8. https://doi.org/10.1109/UIC-ATC.2017.8397566.

Cocca, M., Giordano, D., Mellia, M., Vassio, L., 2019. Free floating electric car sharing design: Data driven optimisation. Pervasive Mob. Comput. 55, 59–75.

Cocca, M., Teixeira, D., Vassio, L., Mellia, M., Almeida, J.M., Da Couto Silva, A.P., 2020. On Car-Sharing Usage Prediction with Open Socio-Demographic Data. Electronics 9, 72.

Cohen, B., Kietzmann, J., 2014. Ride On! Mobility Business Models for the Sharing Economy. Organ. Environ. 27, 279–296.

Correia, G.H.d.A., Antunes, A.P., 2012. Optimization approach to depot location and trip selection in one-way carsharing systems. Transportation Research Part e: Logistics and Transportation Review 48, 233–247.

Costain, C., Ardron, C., Habib, K.N., 2012. Synopsis of users' behaviour of a carsharing program: A case study in Toronto. Transp. Res. A Policy Pract. 46, 421–434.

Cui, S., Ma, X., Zhang, M., Yu, B., Yao, B., 2022. The parallel mobile charging service for free-floating shared electric vehicle clusters. Transportation Research Part E: Logistics and Transportation Review 160, 102652.

Davila Delgado, J.M., Oyedele, L., 2021. Deep learning with small datasets: using autoencoders to address limited datasets in construction management. Appl. Soft Comput. 112, 107836.

de Lorimier, A., El-Geneidy, A.M., 2013. Understanding the Factors Affecting Vehicle Usage and Availability in Carsharing Networks: A Case Study of Communauto Carsharing System from Montréal, Canada. Int. J. Sustain. Transp. 7, 35–51.

Etxandi-Santolaya, M., Canals Casals, L., Corchero, C., 2023. Estimation of electric vehicle battery capacity requirements based on synthetic cycles. Transp. Res. Part D: Transp. Environ. 114, 103545.

Ferrero, F., Perboli, G., Rosano, M., Vesco, A., 2018. Car-sharing services: An annotated review. Sustain. Cities Soc. 37, 501–518.

Figueira, A., Vaz, B., 2022. Survey on Synthetic Data Generation, Evaluation Methods and GANs. Mathematics 10, 2733.

Garrido, S., Borysov, S.S., Pereira, F.C., Rich, J., 2020. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. Transportation Research Part C: Emerging Technologies 120, 102787.

Giordano, D., Vassio, L., Cagliero, L., 2021. A multi-faceted characterization of free-floating car sharing service usage. Transportation Research Part C: Emerging Technologies 125, 102966.

Golalikhani, M., Oliveira, B.B., Carravilla, M.A., Oliveira, J.F., Antunes, A.P., 2021a. Carsharing: A review of academic literature and business practices toward an integrated decision-support framework. Transportation Research Part E: Logistics and Transportation Review 149, 102280.
Golalikhani, M., Oliveira, B.B., Carravilla, M.A., Oliveira, J.F., Pisinger, D., 2021b. Understanding carsharing: A review of managerial practices towards relevant research insights. Res. Transp. Bus. Manag., 100653
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. https://doi.org/10.48550/arXiv.1406.2661.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63, 139–144.
Goodfellow, I., 2017. NIPS 2016 Tutorial: Generative Adversarial Networks. p. 57–pp. http://arxiv.org/pdf/1701.00160v4.
Gudivada, V., Apon, A., Ding, J., 2017. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. International Journal on Advances in Software 10, 1–20.
Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved Training ofWasserstein GANs. Advances in neural information processing systems 30. https://doi.org/10.48550/arXiv.1704.00028.
Hahn, R., Ostertag, F., Lehr, A., Büttgen, M., Benoit, S., 2020. "I like it, but I don't use it": Impact of carsharing business models on usage intentions in the sharing economy. Bus Strat Env 29, 1404–1418.
He, Z., Chen, P., 2021. Shared mobility: Characteristics, impacts, and improvements. Transp. Res. Part D: Transp. Environ. 97, 102960.
He, Z., Zhou, W., 2022. Generation of synthetic full-scale burst test data for corroded pipelines using the tabular generative adversarial network. Eng. Appl. Artif. Intel. 115, 105308.
Hoerler, R., van Dijk, J., Patt, A., Del Duce, A., 2021. Carsharing experience fostering sustainable car purchasing? Investigating car size and powertrain choice. Transp. Res. Part D: Transp. Environ. 96, 102861.
Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., Wang, H.-M., 2017. Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks. p. 5–pp. https://arxiv.org/pdf/1704.00849v3.pdf.
Hu, S., Chen, P., Lin, H., Xie, C., Chen, X., 2018a. Promoting carsharing attractiveness and efficiency: An exploratory analysis. Transp. Res. Part D: Transp. Environ. 65, 229–243.
Hu, S., Lin, H., Xie, K., Chen, X., Shi, H., 2018b. Modeling users' vehicles selection behavior in the urban carsharing program. In: 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 1546–1551.
Hu, S., Chen, P., Xin, F., Xie, C., 2019. Exploring the effect of battery capacity on electric vehicle sharing programs using a simulation approach. Transp. Res. Part D: Transp. Environ. 77, 164–177.
Hu, S., Chen, P., Chen, X., 2021. Do personalized economic incentives work in promoting shared mobility? Examining customer churn using a time-varying Cox model. Transportation Research Part C: Emerging Technologies 128, 103224.
Hu, L., Liu, Y., 2016. Joint design of parking capacities and fleet size for one-way station-based carsharing systems with road congestion constraints. Transp. Res. B Methodol. 93, 268–299.
Illgen, S., Höck, M., 2019. Literature review of the vehicle relocation problem in one-way car sharing networks. Transp. Res. B Methodol. 120, 193–204.
Illgen, S., Höck, M., 2020. Establishing car sharing services in rural areas: a simulation-based fleet operations analysis. Transportation 47, 811–826.
Inan, M.S.K., Hossain, S., Uddin, M.N., 2023. Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor's morphological information. Inf. Med. Unlocked 37, 101171.
Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J., 2021. Crash data augmentation using variational autoencoder. Accid. Anal. Prev. 151, 105950.
Jian, S., Rashidi, T.H., Dixit, V., 2017. An analysis of carsharing vehicle choice and utilization patterns using multiple discrete-continuous extreme value (MDCEV) models. Transp. Res. A Policy Pract. 103, 362–376.
Jordon, J., Wilson, A., van der Schaar, M., 2020. Synthetic Data: Opening the data floodgates to enable faster, more directed development of machine learning methods 9. http://arxiv.org/pdf/2012.04580v1.
Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N., Weller, A., 2022. Synthetic Data – what, why and how? p. 57–pp. http://arxiv.org/pdf/2205.03257v1.
Jorge, D., Correia, G., 2013. Carsharing systems demand estimation and defined operations: a literature review. Eur. J. Transp. Infrastruct. Res. 13 (3), 201–220. https://doi.org/10.18757/ejtir.2013.13.3.2999.
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.
Kim, D., Park, Y., Ko, J., 2019. Factors underlying vehicle ownership reduction among carsharing users: A repeated cross-sectional analysis. Transp. Res. Part D: Transp. Environ. 76, 123–137.
Kingma, D.P., Welling, M., 2013. Auto-Encoding Variational Bayes 14. http://arxiv.org/pdf/1312.6114v10.
Kingma, D.P., Welling, M., 2019. An Introduction to Variational Autoencoders. FNT in Machine Learning 12, 307–392.
Kostic, B., Loft, M.P., Rodrigues, F., Borysov, S.S., 2021. Deep survival modelling for shared mobility. Transportation Research Part C: Emerging Technologies 128, 103213.
Kühl, N., Hirt, R., Baier, L., Schmitz, B., Satzger, G., 2021. How to Conduct Rigorous Supervised Machine Learning in Information Systems Research: The Supervised Machine Learning Report Card. CAIS 48, 589–615.
Kuhn, M., Johnson, K., 2019. Feature Engineering and Selection. Chapman and Hall/CRC.
Lagadic, M., Verloes, A., Louvet, N., 2019. Can carsharing services be profitable? A critical review of established and developing business models. Transp. Policy 77, 68–78.
Lai, M., Hu, Q., Liu, Y., Lang, Z., 2022. A rolling-horizon decision framework for integrating relocation and user flexibility in one-way electric carsharing systems. Transportation Research Part C: Emerging Technologies 144, 103867.
Laporte, G., Meunier, F., Wolfler Calvo, R., 2018. Shared mobility systems: an updated survey. Annals of Operations Research 271, 105–126.
Lei, X., Veeramachaneni, K., 2018. Synthesizing Tabular Data using Generative Adversarial Networks 12. http://arxiv.org/pdf/1811.11264v1.
Lei, X., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K., 2019. Modeling Tabular data using Conditional GAN, 15 pp. http://arxiv.org/pdf/1907.00503v2.
Lei, Z., Qian, X., Ukkusuri, S.V., 2020. Efficient proactive vehicle relocation for on-demand mobility service with recurrent neural networks. Transportation Research Part C: Emerging Technologies 117, 102678.
Li, D.X., 1999. On Default Correlation: A Copula Function Approach. SSRN Journal. https://doi.org/10.2139/ssrn.187289.
Li, M., Zeng, Z., Wang, Y., 2021. An innovative car sharing technological paradigm towards sustainable mobility. J. Clean. Prod. 288, 125626.
Li, L., Zhu, J., Zhang, H., Tan, H., Du, B., Ran, B., 2020. Coupled application of generative adversarial networks and conventional neural networks for travel mode detection using GPS data. Transp. Res. A Policy Pract. 136, 282–292.
Lin, Z., Khetan, A., Fanti, G., Oh, S., 2020. PacGAN: The Power of Two Samples in Generative Adversarial Networks. IEEE J. Sel. Areas Inf. Theory 1, 324–335.
Little, R.J.A., 1993. Statistical Analysis of Masked Data. J. Off. Stat. 9, 407–426.
Lu, Y., Wang, K., Yuan, B., 2022. The vehicle relocation problem with operation teams in one-way carsharing systems. Int. J. Prod. Res. 60, 3829–3843.
Luna, T.F., Uriona-Maldonado, M., Silva, M.E., Vaz, C.R., 2020. The influence of e-carsharing schemes on electric vehicle adoption and carbon emissions: An emerging economy study. Transp. Res. Part D: Transp. Environ. 79, 102226.
Ma, Y., Miao, R., Chen, Z., Zhang, B., Bao, L., 2022. An interpretable analytic framework of the relationship between carsharing station development patterns and built environment for sustainable urban transportation. J. Clean. Prod. 377, 134445.
Ma, T.-Y., Xie, S., 2021. Optimal fast charging station locations for electric ridesharing with vehicle-charging station assignment. Transp. Res. Part D: Transp. Environ. 90, 102682.
Massey Jr., F.J., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. J. Am. Stat. Assoc. 46, 68–78.

Meng, Z., Li, E.Y., Qiu, R., 2020. Environmental sustainability with free-floating carsharing services: An on-demand refueling recommendation system for Car2go in Seattle. Technol. Forecast. Soc. Chang. 152, 119893.
Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets 7. http://arxiv.org/pdf/1411.1784v1.
Molnar, G., Correia, G.H.d.A., 2019. Long-term vehicle reservations in one-way free-floating carsharing systems: A variable quality of service model. Transportation Research Part C: Emerging Technologies 98, 298–322.
Mounce, R., Nelson, J.D., 2019. On the potential for one-way electric vehicle car-sharing in future mobility systems. Transp. Res. A Policy Pract. 120, 17–30.
Münzel, K., Boon, W., Frenken, K., Vaskelainen, T., 2018. Carsharing business models in Germany: characteristics, success and future prospects. Inf Syst E-Bus Manage 16, 271–291.
Münzel, K., Piscicelli, L., Boon, W., Frenken, K., 2019. Different business models – different users? Uncovering the motives and characteristics of business-to-consumer and peer-to-peer carsharing adopters in The Netherlands. Transp. Res. Part D: Transp. Environ. 73, 276–306.
Nansubuga, B., Kowalkowski, C., 2021. Carsharing: a systematic literature review and research agenda. JOSM 32, 55–91.
Niels, T., Bogenberger, K., 2017. Booking Behavior of Free-Floating Carsharing Users. Transp. Res. Rec. 2650, 123–132.
Nijland, H., van Meerkerk, J., 2017. Mobility and environmental impacts of car sharing in the Netherlands. Environ. Innov. Soc. Trans. 23, 84–91.
Nikolenko, S.I., 2021. Synthetic Data for Deep Learning, 1st ed. Springer International Publishing; Imprint Springer, Cham, p. 348.
Nourinejad, M., Roorda, M.J., 2015. Carsharing operations policies: a comparison between one-way and two-way systems. Transportation 42, 497–518.
Open Mobility Foundation, 2023. Mobility Data Specification. https://github.com/openmobilityfoundation/mobility-data-specification Accessed 22 June 2023.
OpenWeather, 2023. Historical weather data for Augsburg, Germany. https://openweathermap.org/history-bulk Accessed 10 May 2023.
Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y., 2018. Data Synthesis based on Generative Adversarial Networks. Proc. VLDB Endow. 11, 1071–1083.
Patki, N., Wedge, R., Veeramachaneni, K., 2016. The Synthetic Data Vault, in: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada. 17.10.2016 - 19.10.2016. IEEE, pp. 399–410. https://doi.org/10.1109/DSAA.2016.49.
Perboli, G., Ferrero, F., Musso, S., Vesco, A., 2018. Business models and tariff simulation in car-sharing services. Transp. Res. A Policy Pract. 115, 32–48.
Prinz, C., Lichtenberg, S., Herrenkind, B., Brendel, A.B., Kolbe, L., 2020a. Adapting User-Based Vehicle Relocation for E-Carsharing. 15th International Conference on Wirtschaftsinformatik, 1490–1505.
Prinz, C., Hellmeier, M., Willnat, M., Harnischmacher, C., Kolbe, L., 2022. Increasing the Business Value Of Free-Floating Carsharing Fleets By Applying Machine-Learning Based Relocations. Proceedings of the 30th European Conference on Information Systems (ECIS). https://aisel.aisnet.org/ecis2022_rp/70.
Prinz, C., Lichtenberg, S., Willnat, M., 2020b. CASSI: Design of a Simulation Environment for Vehicle Relocation in Carsharing. Proceedings of the 28th European Conference on Information Systems (ECIS). https://aisel.aisnet.org/ecis2020_rp/103.
Ratliff, L.J., Burden, S.A., Sastry, S.S., 2013. Characterization and computation of local Nash equilibria in continuous games, in: 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello, IL. 02.10.2013 - 04.10.2013. IEEE, pp. 917–924. https://doi.org/10.1109/Allerton.2013.6736623.
Ren, C., An, L., Gu, Z., Wang, Y., Gao, Y., 2020. Rebalancing the car-sharing system with reinforcement learning. World Wide Web 23, 2491–2511.
Rotaris, L., Danielis, R., 2018. The role for carsharing in medium to small-sized towns and in less-densely populated rural areas. Transp. Res. A Policy Pract. 115, 49–62.
Rubin, D.B., 1993. Statistical Disclosure Limitation. J. Off. Stat. 9, 461–468.
SAE International, 2021. Taxonomy of On-Demand and Shared Mobility: Ground, Aviation, and Marine. https://www.sae.org/standards/content/ja3163_202106/ Accessed 21 June 2023.
Schmöller, S., Bogenberger, K., 2020. Carsharing: An overview on what we know. In: Demand for Emerging Transportation Systems. Elsevier, pp. 211–226. https://doi.org/10.1016/B978-0-12-815018-4.00011-5.
Schmöller, S., Weikl, S., Müller, J., Bogenberger, K., 2015. Empirical analysis of free-floating carsharing usage: The Munich and Berlin case. Transportation Research Part C: Emerging Technologies 56, 34–51.
Schroer, K., Ketter, W., Lee, T.Y., Gupta, A., Kahlen, M., 2022. Data-Driven Competitor-Aware Positioning in On-Demand Vehicle Rental Networks. Transp. Sci. 56, 182–200.
Sengupta, S., Basak, S., Saikia, P., Paul, S., Tsalavoutis, V., Atiah, F., Ravi, V., Peters, A., 2020. A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl.-Based Syst. 194, 105596.
Shaheen, S., Cohen, A., 2020. Innovative Mobility: Carsharing Outlook; Carsharing Market Overview, Analysis, and Trends - Spring 2020, 7 pp. https://escholarship.org/uc/item/9jh432pm.
Shaheen, S.A., Cohen, A.P., Roberts, J.D., 2006. Carsharing in North America: Market Growth, Current Developments, and Future Potential. Transp. Res. Rec. 1986, 116–124.
Shaheen, S., Martin, E., Totte, H., 2020. Zero-emission vehicle exposure within U.S. carsharing fleets and impacts on sentiment toward electric-drive vehicles. Transp. Policy 85, A23–A32.
Shen, Z.-J.-M., Feng, B., Mao, C., Ran, L., 2019. Optimization models for electric vehicle service operations: A literature review. Transp. Res. B Methodol. 128, 462–477.
Shrestha, Y.R., Krishna, V., von Krogh, G., 2021. Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. J. Bus. Res. 123, 588–603.
Snoke, J., Raab, G.M., Nowok, B., Dibben, C., Slavkovic, A., 2018. General and specific utility measures for synthetic data. J. R. Stat. Soc. A 181, 663–688.
Stone, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. J. Roy. Stat. Soc.: Ser. B (Methodol.) 36, 111–133.
Tanaka, F.H.K.d.S., Aranha, C., 2019. Data Augmentation Using GANs 16. http://arxiv.org/pdf/1904.09135v1.
van Dun, C., Moder, L., Kratsch, W., Röglinger, M., 2022. ProcessGAN: Supporting the creation of business process improvement ideas through generative machine learning. Decis. Support Syst., 113880
Vanheusden, W., van Dalen, J., Mingardo, G., 2022. Governance and business policy impact on carsharing diffusion in European cities. Transp. Res. Part D: Transp. Environ. 108, 103312.
Vélez, A.M.A., 2023. Economic impacts, carbon footprint and rebound effects of car sharing: Scenario analysis assessing business-to-consumer and peer-to-peer car sharing. Sustain. Prod. Consumpt. 35, 238–249.
Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.-Y., 2017. Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sinica 4, 588–598.
Wang, N., Guo, J., Liu, X., Fang, T., 2020. A service demand forecasting model for one-way electric car-sharing systems combining long short-term memory networks with Granger causality test. J. Clean. Prod. 244, 118812.
Wang, T., Hu, S., Jiang, Y., 2021. Predicting shared-car use and examining nonlinear effects using gradient boosting regression trees. Int. J. Sustain. Transp. 15, 893–907.
Wappelhorst, S., Sauer, M., Hinkeldein, D., Bocherding, A., Glaß, T., 2014. Potential of Electric Carsharing in Urban and Rural Areas. Transp. Res. Procedia 4, 374–386.
Weikl, S., Bogenberger, K., 2015. A practice-ready relocation model for free-floating carsharing systems with electric vehicles – Mesoscopic approach and field trial results. Transportation Research Part C: Emerging Technologies 57, 206–223.
Wielinski, G., Trépanier, M., Morency, C., 2019. Exploring Service Usage and Activity Space Evolution in a Free-Floating Carsharing Service. Transp. Res. Rec. 2673, 36–49.
Willing, C., Klemmer, K., Brandt, T., Neumann, D., 2017. Moving in time and space – Location intelligence for carsharing decision support. Decis. Support Syst. 99, 75–85.

Wong, M., Farooq, B., 2020. A bi-partite generative model framework for analyzing and simulating large scale multiple discrete-continuous travel behaviour data. Transportation Research Part C: Emerging Technologies 110, 247–268.
Xu, M., Meng, Q., 2019. Fleet sizing for one-way electric carsharing services considering dynamic vehicle relocation and nonlinear charging profile. Transp. Res. B Methodol. 128, 23–49.
Xu, M., Wu, T., Tan, Z., 2021. Electric vehicle fleet size for carsharing services considering on-demand charging strategy and battery degradation. Transportation Research Part C: Emerging Technologies 127, 103146.
Yang, S., Wu, J., Sun, H., Qu, Y., Wang, D.Z., 2022. Integrated optimization of pricing and relocation in the competitive carsharing market: A multi-leader-follower game model. Transportation Research Part C: Emerging Technologies 138, 103613.
Yao, R., Bekhor, S., 2022. A variational autoencoder approach for choice set generation and implicit perception of alternatives in choice modeling. Transp. Res. B Methodol. 158, 273–294.
Yao, Z., Gendreau, M., Li, M., Ran, L., Wang, Z., 2022. Service operations of electric vehicle carsharing systems from the perspectives of supply and demand: A literature review. Transportation Research Part C: Emerging Technologies 140, 103702.
Yoon, T., Cherry, C.R., Ryerson, M.S., Bell, J.E., 2019. Carsharing demand estimation and fleet simulation with EV adoption. J. Clean. Prod. 206, 1051–1058.
Zhang, K., Jia, N., Zheng, L., Liu, Z., 2019. A novel generative adversarial network for estimation of trip travel time distribution with trajectory data. Transportation Research Part C: Emerging Technologies 108, 223–244.
Zhang, C., Schmöcker, J.-D., Kuwahara, M., Nakamura, T., Uno, N., 2020. A diffusion model for estimating adoption patterns of a one-way carsharing system in its initial years. Transp. Res. A Policy Pract. 136, 135–150.
Zhou, L., Pan, S., Wang, J., Vasilakos, A.V., 2017. Machine learning on big data: Opportunities and challenges. Neurocomputing 237, 350–361.
Zhu, X., Li, J., Liu, Z., Yang, F., 2017. Location deployment of depots and resource relocation for connected car-sharing systems through mobile edge computing. International Journal of Distributed Sensor Networks 13. https://doi.org/10.1177/1550147717711621.
Zhu, H., Luo, Y., Liu, Q., Fan, H., Song, T., Yu, C.W., Du, B., 2019. Multistep flow prediction on car-sharing systems: a multi-graph convolutional neural network with attention mechanism. Int. J. Soft. Eng. Knowl. Eng. 29, 1727–1740.