Data Science for Transportation (2023) 5:14
https://doi.org/10.1007/s42421-023-00079-6

RESEARCH

Synthetic Generation of Trip Data: The Case of Smart Card

Minh Kieu[1] · **Iris Brighid Meredith**[2] · **Andrea Raith**[2]

Received: 25 August 2022 / Revised: 30 May 2023 / Accepted: 20 July 2023 / Published online: 31 July 2023
© The Author(s) 2023

Abstract
While individual data are key for epidemiology, social simulation, economics, and various other fields, data owners are increasingly required to protect the personally identifiable information from data. Simple data de-identification or 'data masking' measures are limited, because they both reduce the utility of the dataset and are not sufficient to protect individual confidentiality. This paper provides detail on the creation of a synthetic trip data in transportation, with the Smart Card data as the case study. It discusses and compares two machine learning methods, a Generative Adversarial Network and a Bayesian Network for modelling and generating this dataset. The synthetic data retain important utility of the real dataset, e.g., the origin, destination, and time of travel, while each data point does not represent a real trip in the original dataset. The synthetic dataset can be used in various applications, including microsimulation of public transport systems, analysing travel behaviours, model predictive control of transit flows, or evaluation of transport policies.

Keywords Synthetic data · Generative Adversarial Network · Bayesian Network · Smart Card data

Introduction

The acquisition and dissemination of individual data are key for research in many disciplines, including epidemiology, social simulation, economics, and engineering. Wide access to individual data has unprecedented benefits for the analysis and modelling of heterogeneous individual behaviour and is highly valuable when supporting decision-making. For instance, during the COVID-19 pandemic, individual data could be used to develop contact tracing and social distancing strategies to limit the spread of the disease and to recover from the pandemic.

However, governments and individuals across the globe have increasingly become concerned over privacy and the exchange of personal data. Data owners such as statistical and operational agencies that disseminate the data are often ethically and legally required to protect the personally identifiable information of individuals. Agencies often implement simple data de-identification measures such as removing obvious identifiers such as names and addresses. In the case of Smart Card data in public transport (the data of the payment cards that often store the tag-on and tag-off locations and times of individual transit riders), data owners often perform a 'data masking' measure, e.g., placing a unique hashed identification number over the original Smart Card ID. Simple data de-identification measures, such as data-masking, top-coding, adding noise, or random data swapping, are not sufficient to protect individual confidentiality (Drechsler and Reiter 2011). In the case of Smart Card data, intruders may observe the same Smart Card user over a period of multiple days and locate an individual's home location and personal travel patterns and trajectories. This poses a major risk of privacy breaches for vulnerable individuals, such as senior or child cardholders, members of ethnic or religious minorities at high risk of hate crimes, and women and members of the LGBTQ+ community. This risk thus prevents the wider dissemination of personal data such as Smart Card data to the wider research community and limits the impacts of research on policy making. On the other hand, as the risks of personal data disclosure increase, the alterations made by data owners using simple data de-identification techniques may impact the usefulness of the released data.

To address the limitations of standard de-identification measures, literature has offered various approaches aiming

✉ Minh Kieu
minh.kieu@auckland.ac.nz

[1] Department of Civil and Environmental Engineering, University of Auckland, Auckland 1010, New Zealand

[2] Department of Engineering Science, University of Auckland, Auckland 1010, New Zealand

at generating partially synthetic or fully synthetic data from real data. The idea is to analyse and model the real data to preserve the ability to query from that data, but not to reveal raw individual data points. In other words, data synthesis aims to retain the distributions in the data, e.g., the mean, standard deviation, and potentially even the probability distribution of the data, but each data sample does not represent a real person in the raw data. Synthetic data enable public dissemination of the data while protecting individual privacy and preserving data utility. With higher quality synthetic data, analysts can develop meaningful and relevant research that can contribute to decision-making. Data owners, who are generally policy makers, can also benefit from access to cutting-edge models and synthesis methods that can be directly implemented on the real data.

While synthetic data generation has attracted great interest and proved effective for images (Karras et al. 2020), music (Briot et al. 2020), and texts (McKeown 1992), the existing studies in transport synthetic data are still currently limited. Most existing works focus on creation of synthetic population for activity-based travel models (Axhausen and Gärling 1992), where various methods such as Iterative Proportion Fitting (Ruschendorf 1995), Maximum Cross-Entropy (Guo Jessica and Bhat Chandra 2007), and Bayesian Network (Sun and Erath 2015) have been developed for the purpose. More recent studies have focused on ultilising more data-driven methods to discover and model hidden mobility patterns, such as Input–Output Hidden Markov models (Yin et al. 2017), semi-Markov models (Samiul and Ukkusuri Satish 2017), and Generative Adversarial Networks (Badu-Marfo et al. 2020).

The main objective of this paper diverges from existing research efforts as it seeks to generate a synthetic version of a large transport dataset that preserves the format and spatio-temporal mobility patterns of individuals, while ensuring that each data point does not represent any individual person. Specifically, we focus on Smart Card data, which has become the prevalent standard in modern public transport systems. Although the availability of Smart Card data has opened up new avenues for research in intelligent transport systems, such as the analysis of travel behaviours (Kieu et al. 2015), identification of trip purposes (Lee and Hickman 2014), and transfer intentions (Kieu et al. 2017), its potential has not been fully realised by the research community, while policy makers who have access to the raw data remain largely unaware of cutting-edge research possibilities. This paper proposes a framework that connects Smart Card data owners to a broader research community through the creation of synthetic data. Additionally, this paper offers a superior option for public data dissemination to public transport agencies, research centres, local councils, and other Smart Card data owners.

This paper addresses three key scientific challenges related to Smart Card data: non-Gaussian distributions, mixed data types, and highly imbalanced and multi-modal data. Real-world Smart Card data are often non-Gaussian, and may include both continuous values (such as tag-on and tag-off timestamps) and discrete values (such as Route, Direction, or Zones). Additionally, the data are frequently highly imbalanced, with certain discrete fields such as Zones or Routes being much more popular than others, and the continuous fields such as tag-on and tag-off likely exhibiting bi-modal distributions reflecting the two peak periods. The scientific contributions of this paper are twofold:

- We apply a Generative Adversarial Network and a Bayesian Network to model and generate synthetic Smart Card data
- We compare and contrast the two methods mentioned above, discussing the advantages and disadvantages of each for the data synthesis problem.

The remainder of this manuscript is structured as follows. Section "Related Works" reviews the literature on synthetic data generation for sequential data. Section "Methodology" discusses three important synthesis methods applied in this paper and compares and contrasts them in the context of synthetic data generation. Section "Numerical Experiments" discusses the application of these synthesis methods to Smart Card data from South East Queensland, Australia's public transport network and compares both the modelling process and synthesised data from the methods. Finally, Sect. "Discussion and Conclusion" concludes this study and suggests several directions for future research.

Related Works

Classical Data De-identification Techniques

Classical data de-identification techniques often aim to create a version of the real data that does not contain or only contains synthetic information about individuals. Common techniques often focus on altering the raw data using techniques ranging from simply removing sensitive attributes (e.g., ID, name, age, and income) to aggregating geography, swapping data across records, and adding noise to the data (Willenborg and Waal 2001). Mendes and Vilela (2017) provides a survey of the most common privacy-preserving data synthesis techniques, which can be classified into four main classes:

- Generalisation: replacement of values with more general ones, e.g., numerical values being generalised as an interval

- Suppression: removal of sensitive attributes, e.g., name or ID
- Anatomisation: de-association of sensitive attributes to multiple tables, making re-identification more difficult
- Perturbation: replacement of the raw data with synthetic values with identical information or swapping data between attributes.

However, recent research examples show that these classical anonymisation techniques fail for 'big data' (Lu et al. 2014; Shrivastva et al. 2014) for two reasons. First, they often protect privacy by reducing the utility of the data (Purdam and Elliot 2007). Second, even the strictest classical data anonymisation techniques may fail to preserve confidentiality from a nefarious person who is reasonably competent in employing investigative techniques (Paul 2009). Sweeney (2001) showed that 97% of the names and addresses on a voting list can be identified using only ZIP codes and date of birth.

The Generation of Synthetic Transport Data

To address the shortcomings of classical data anonymisation techniques, generative models often aim to learn probabilistic distributions in the original data and then generate completely *synthetic* versions of the data. Unlike classical data anonymisation techniques, which mainly focus on releasing altered versions of the raw data, generative models aim to learn the probability distributions in the real data to draw new and synthetic samples that retain the distributions and correlations, but do not contain any record of real individuals.

Synthetic data, in general, should have the same format and statistical distribution as in the real data. This is to ensure that data users can get the same insights from the synthetic data as from the real data. Conversely, to maximise their benefits, synthetic data should be able to be used in place of the real data in data-driven models, or in other words, models trained on the synthetic data should be able to be applied directly to the real data. This allows data owners to implement data-driven models trained on synthetic data directly to the real data. To this end, there is a vast and recent literature on the use of generative deep learning methods, such as Generative Adversarial Networks (GAN) (Goodfellow et al. 2014), generative stochastic networks (Bengio et al. 2014), Bayesian Networks (Deeva et al. 2020), Hidden Markov Models (Ismaïl et al. 2016), and variational autoencoders (VAEs) (Kingma and Welling 2014). In recent years, these methods have been extended to cater various specific data and have achieved notable successes with images (Karras et al. 2020), music (Briot et al. 2020), texts (McKeown 1992), and tabular data (Xu et al. 2019). However, there are generally fewer combinations of letters and musical notes than spatial locations and timestamps in travel activities.

The generation of synthetic travel activities is related to a major class of transport models: activity-based models (Rasouli and Timmermans 2014; Axhausen and Gärling 1992), as they also aim to reproduce realistic individual travel activities. There are a rich literature of studies focusing on synthetic population for activity-based travel models, as these models often require specific data of location and timestamps of individual travels. Classical methods in this topic include the *deterministic* approaches such as Iterative Proportion Fitting (Ruschendorf 1995) and Maximum Cross-Entropy (Guo Jessica and Bhat Chandra 2007), as well as the *probabilistic* methods such as Bayesian Network (Sun and Erath 2015).

Recent studies have proposed more data-driven methods that can directly learn from data and produce synthetic samples of similar data format and statistical distributions. To this end, there are Markov-based studies that aims to model human travel activities as probabilistic and stochastic changes of discrete states. Samiul and Ukkusuri Satish (2017) aims to reconstruct activity–location sequences of travel activities from online social media data with imcompleted information. A semi-Markov modelling approach was proposed, along with a particle-based Markov chain Monte Carlo sampler to perform parameters inference for the model. Jiang et al. (2016) proposed TimeGeo, a mechanistic modelling framework that aims to synthetically generate travel activities with high resolution. TimeGeo integrates a Markov chain to generate temporal travel patterns with a rank-based exploration and preferential return (r-EPR) mechanism to generate spatial patterns. Yin et al. (2017) proposes an Input–Output Hidden Markov models that utilises contextual information to address the limitations of the standard Hidden Markov models due to its uniform transition and emission probabilities. Also, based on a Markov chain, Pappalardo and Simini (2018) extend the work by Jiang et al. (2016) by proposing a mechanistic model that is more data-driven and parameter-free. The proposed model in Pappalardo and Simini (2018) aims to generate individual trajectories in fine spatio-temporal resolution. Their proposed model generates a mobility diary by learning from the data, and then generates a mobility trajectory based on the concept of preferential exploration and preferential return. Although mechanistic models, such as Pappalardo and Simini (2018), Jiang et al. (2016) and Yin et al. (2017), are often interpretable, they can lack the complexity to model complex mobility patterns of a large number of individuals (Choi et al. 2021). For this reason, deep-learning-based generative mobility modelling has recently become the emerging approach in the literature. Badu-Marfo et al. (2020) developed a Generative Adversarial Network (GAN) to synthesise daily travel activities. The GAN model was trained in a differentially private manner to guarantee the privacy of the synthesis seed sample

members. Choi et al. (2021) proposed a generative adversarial imitation learning framework based on GAN to generate trajectory data in urban transport. Individual decisions on the road network were modelled as a partially observable Markov decision process.

The Generation of Synthetic Smart Card Data

This paper aims to create a simulated edition of a vast transportation dataset that retains the structure and temporal movement behaviours of individuals, while guaranteeing that no individual's data point is identifiable. The case study is the generation of synthetic Smart Card data in public transport. Smart Card data have recently been utilised for various purposes, but the data have only been accessible to a very limited number of researchers. If a wider research community could access the (synthetic) Smart Card data, we might be able to see many more innovations to support policymakers in making informed decisions in public transport. For instance, exploratory data analysis or simulation models can be developed on the synthetic dataset and then applied on the real Smart Card dataset by data owners.

To this end, Bouman et al. (2017) proposed an approach for the generation of synthetic Smart Card data. However, the proposed Markov Chain was not calibrated against real data, making it impossible to evaluate the similarity of the synthetic data to the real data. In this paper, we focus on a more data-driven approach to generate synthetic Smart Card data that are as similar as possible to real data. To this end, we identify several challenges in the development of Smart Card data synthesisers:

- Non-Gaussian distributions: the real-world Smart Card data are unlikely to be Gaussian, which may lead to the vanishing gradient problem during data normalisation (Xu et al. 2019)
- Mixed data types: Smart Card data often consist of both continuous values (e.g., tag-on and tag-off timestamps) and discrete values (e.g., Route, Direction, or Zones).
- Highly imbalanced and multi-modal data: Smart Card data are often highly imbalanced, especially on some discrete fields, such as Zones or Routes, as some are more popular than others. The continuous fields such as tag-on and tag-off are likely bi-modal, reflecting the two peak periods.

Methodology

Data Description

This paper uses real, individual Smart Card data transactions from Brisbane, Australia as the case study to show the effectiveness of the proposed methods. Smart Card has been widely used as public transport fare card around the world. A version of Smart Card is now available in many major cities, such as the Oyster Card in London (UK), Opal Card in Sydney (Australia), or Hop Card in Auckland (New Zealand). This paper uses raw Smart Card data of a 3-month period of bus trips associated with the South East Busways (Translink 2019). The data consist of over 1 million Smart Card transactions from July to October 2015. The dataset is provided by Translink, the transit authority of South East Queensland, Australia. We have processes the data to remove the unnecessary fields for this project, and to remove the individual Smart Card identifier. Each processed data record contains the following fields:

- Trip ID: Unique identifier for each bus trip
- Route: Bus route (e.g., 555)
- Direction: Direction of travel (Inbound or Outbound)
- Origin zone: The zone ID of trip origin (1–23)
- Destination zone: The zone ID of trip destination (1–23)
- Month: Month of the trip (July–Oct)
- Weekday: Monday–Friday
- Tag-on: The time a Smart Card user enters the bus (e.g., 08:23 AM)
- Tag-off: The time the Smart Card user leaves the bus (e.g., 08:47 AM).

The Origin and Destination (O &D) zone utilised in this study adheres to the pre-2016 policy of Translink, featuring a roughly concentric circular system of 23 zones. The first zone encompasses the central business district of Brisbane, while the 23rd zone represents the farthest location from Brisbane, namely Noosa. Figure 1 illustrates the zoning system in the dataset (Wallner et al. 2018).

Generative Adversarial Network (GAN)

Generative Adversarial Networks (GANs) are generative models in deep learning that aim to discover the patterns in input data and then generate new data observations that are very similar to the original dataset. The core idea of GANs is the use of two sub-models: a generator model that generates new observations, and a discriminator model that classifies the generated observations as either real or fake data. The two sub-models are trained subsequently in a game theoretic zero-sum game, until the discriminator cannot differentiate the generated from the real observations for half of the time, which means that the generator is capable of generating valid observations. More details on GAN can be found in the original paper by Goodfellow et al. (2014).

GANs and their numerous extensions have been emerging in recent literature for various generative purposes, such as generating images (Karras et al. 2020) or health data (Yale

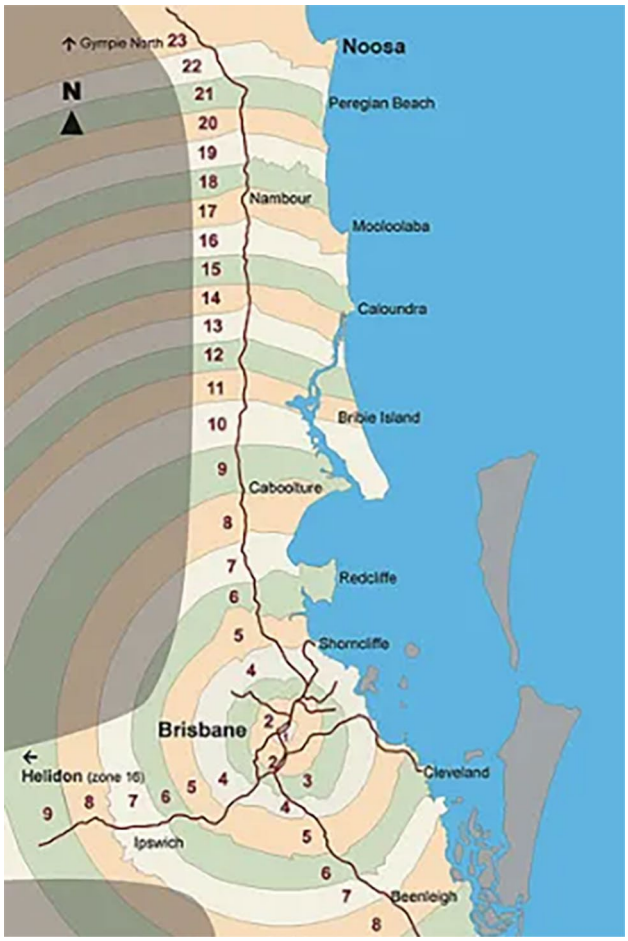


Fig. 1 South East Queensland Translink's zones

et al. 2020), due to their performance an versatility in modelling and generating synthetic data. We are particularly interested in Tabular Conditional GAN (CTGAN), an extension of GAN that excels in modelling and generating mixed tabular data of continuous and discrete variables (Xu et al. 2019). CTGAN excels in modelling and generating mixed tabular data of continuous and discrete variables, similar to our Smart Card data. CTGAN has been proven to outperform many other data generative methods in the original paper (Xu et al. 2019) and several specific applications, such as generating synthetic insurance data (Kuo 2019) or diseases data in agriculture (Ahmed et al. 2021).

Compared to the original GAN, CTGAN introduces two new techniques enhancing the quality of data generation for tabular data such as our Smart Card data: mode-specific normalisation and conditional sample training. Mode-specific normalisation enhances the modelling of multi-modal distributions in numeric variables, and conditional sample training helps with the imbalanced level frequencies in categorical variables in the data. Both of these are necessary for our numeric variables (both tag-on and tag-off have multiple peaks corresponding to the morning and afternoon peak periods) and categorical variables (variables such as Origin zone and Destination zone are imbalanced, with the majority of people travelling to and from the Zone 1—the city centre in our data).

Bayesian Networks (BN)

In addition to the Generative Adversarial Network approach, a statistical approach using Bayesian Network models was developed for comparison and to explore another avenue for synthetic data generation. A Bayesian Network (BN) (Sun and Erath 2015) is a directed acyclic graph (DAG) with a variable and a conditional probability associated with each node: the distribution for each variable is conditioned on the variables upstream of it in the DAG. A joint probability distribution for the variables can then be fitted on the graph and sampled from, generating a synthetic sample. BN and GAN represent two very different approaches to synthetic data generation. The BN tackles the problem as a statistician might: an appropriate, well-understood model is fitted to the data, parameters are fitted to the model, and the model can then be studied or used to generate synthetic data.

One distinction of great importance in the current context is that the BN approach is much more transparent than the GAN approach. Shape learning gives an easily understood dependency structure, parameter fitting using maximum-likelihood estimation is a well-understood statistical method, and forward sampling is a robust sampling method. Between all these criteria, it is easy to deduce why the model does what it does. The GAN method, by contrast, is highly opaque: it is difficult to understand how the network learns and makes its decisions. This is important in a context where it is necessary to be able to prove that the data are anonymous, does not correspond to any particular individual, and does not expose aggregate information that could be problematic. On the other hand, Bayesian approaches are much less easy to generalise: they tend to require significant understanding of the underlying dataset and its statistical qualities, they often require the data to be preprocessed to allow a good model to be fitted, and they do not handle hybrid continuous–discrete datasets very well, usually requiring continuous data to be discretised to allow for good performance. Finally, temporality can be a challenge for Bayesian Networks: while Dynamic and Continuous-Time Bayesian Network (Yan-Feng et al. 2015) formulations do exist to handle temporal data, they are fairly new techniques and require the temporal dependency to be structured in a fairly restrictive manner.

The modelling process for data generation using Bayesian approaches takes place in three broad stages. In the first stage (shape learning problem), the shape of the dependency graph for the data is learned and plotted using one of a

number of possible algorithms. The model's parameters are then fitted using an Expectation–Maximisation approach, and finally, a synthetic data sample can be generated by forward sampling from the fitted model. As shape learning is the first stage of the modelling and generation process, it is also discussed first here.

For simple cases (those with few variables in the dataset) of the shape learning problem, an exhaustive search of all possible dependency graph structures scored by a criterion such as the Bayes Information Criterion can work well. However, as the number of possible DAGs on n labelled vertices follows the sequence 1, 1, 3, 25, 543, 29281, 3781503..., an exhaustive search becomes prohibitively expensive after about four or five variables in the dataset, meaning that more sophisticated search methods are needed.

The Hill Climbing Algorithm is a greedy algorithm that begins with the hypothesis of a disconnected network, and then adds one arc at a time on the basis of whether the addition increases the scoring criterion or not. This runs fast and uses resources efficiently. The major issue with the Hill Climbing Algorithm is that it terminates when it hits a local minimum, and as the shape learning problem is almost certainly not convex, the chances of the algorithm generating an optimal solution for anything except the simplest problems are small. Still, the algorithm has the advantage of being fast, especially by comparison with a constraint-based search, and it often gives remarkably good results.

An alternative approach to Hill Climbing is the Constraint-Based Search approach. This algorithm relies on identifying variables in the data that are independent of each other using the chi-squared test for conditional independence. Having identified the set of conditionally independent variables, a subset of all possible graphs fitting the some constraints. As the search space is considerably reduced by the constraints, the space can be efficiently searched without consuming excessive computational resources. The Constraint-Based search creates graph structures that are in general more connected than the Hill Climbing approach and takes longer to run, but it is less likely to get stuck on a local optimum in the way that the Hill Climbing algorithm is. In this paper, shape learning on the dataset was performed using both the Hill Climbing and Constraint-Based search approaches.

Having fitted a DAG, parameters can then be fitted to the generated DAG. In the process of parameter learning, the goal is to estimate the values of the conditional probability distributions (CPDs). The most popular parameter estimation method is maximum likelihood (ML), which uses relative frequencies or simply the frequency of occurrence for each state of the variable. Although the ML estimator is simple to use, it suffers from overfitting when the observed data does not represent the underlying distribution. Even with a large sample size of training data (such as in our case study with the Smart Card data), the conditional state counts for each parent configuration cause significant fragmentation, particularly when a variable has multiple parents with numerous states. As a result, ML estimation is fragile and unstable for learning Bayesian network parameters.

To overcome this, Bayesian parameter estimation is a useful approach for mitigating the overfitting associated with ML estimation. The Bayesian parameter estimator employs prior CPDs to express our pre-existing beliefs about variables before observing the data. The estimator updates these priors by incorporating state counts from the observed data. These priors can be considered as pseudo-state counts that are added to the actual counts before normalisation.

Finally, having trained a generative model, a synthetic sample can be created by forward sampling from the joint distribution.

Evaluation of Synthetic Data

Here, we propose a tri-step evaluation procedure for the quality of synthetic data generated by BN and GAN, as follows.

The first evaluation step entails analysing the probabilistic spatial–temporal distributions of the synthetic data in comparison to real Smart Card data. This involves examining distributions that relate directly to the variables of the synthetic and real data, as well as those that result from multiple variables (such as travel time as the difference between $Tag-on$ and $Tag-off$). Additionally, statistical evaluations of these distributions are conducted using the two-sample Kolmogorov–Smirnov (KS) and Chi-squared (CS) tests. We also introduce a new spatial distribution distance index to explore the performance of generated origin–destination pairs and the real data.

The existing statistical tests and distance indexes fall short in adequately comparing mixed data. To address this, the second evaluation step employs optimal transport theory, specifically the Wasserstein metric, which can measure the distance between probability distributions, even if the data are not on the same probability space. The Wasserstein metric computes the cost of moving one distribution to match another, with smaller distances indicating closer distributions.

Finally, to comply with privacy regulations and prevent exposing real data to end-users, synthetic data samples are generated using the BN and GAN methods for the final evaluation step. The training data are then examined for duplicates, and this process is iterated using a bootstrapping procedure.

Numerical Experiments

We implement both GAN and BN on the Smart Card data from South East Queensland, Australia. All the developed codes, fitted models, and generated synthetic Smart Card data (excluding the original Smart Card data) are available at an Open Science Framework repository at osf.io/2xep5/.

Bayesian Network Implementations

We first convert the timestamps for *Tag_on* and *Tag_off* into minutes from midnight to make them continuous variables ranging from zero to 1440 min. Because BN learns categorical/discrete variables better, we convert the continuous variables (*Tag_on* and *Tag_off*) to categorical only in a dataset that we use specifically for BN, while keeping them as continuous for the GAN. Experiments show that a round to the nearest 5 min interval works best for BN (e.g., if *Tag_on* equals 451.4 min, we will round the value to 450 min).

In BNs, a graphical model known as a directed acyclic graph (DAG) is used to depict probabilistic connections between variables. The edges between nodes in a BN reflect conditional dependencies between the variables, and each node in the DAG represents a random variable. A directed edge from node A to node B shows that variable A directly influences variable B. Shape learning for the BN was performed with bootstrap aggregation ($n' = 50{,}000$). Constraint-Based learning and Hill Climbing algorithms for shape learning were both evaluated, but while both learned similar dependency structures the Constraint-Based learning algorithm tended to over-fit the dependency structure, leading to a less parsimonious model. The final dependency structure chosen for the Bayesian Network was thus the one learned by the Hill Climbing Algorithm, which is shown in Fig. 2. Each node in the graph corresponds to a random variable in the dataset, and each directed edge between nodes represents a conditional probability distribution.

We can read the arrows in Fig. 2 as directed conditional probability distribution, which can be represented mathematically using Bayes' rule. For instance, the arrow from *Tag_on* to *Tag_off* can be mathematically written as

$$\mathbb{P}(Tag_on|Tag_off) = \frac{\mathbb{P}(Tag_on, Tag_off)}{\mathbb{P}(Tag_on)}, \tag{1}$$

where $\mathbb{P}(Tag_on, Tag_off)$ is the joint probability distribution of *Tag_on* to *Tag_off*, and $\mathbb{P}(Tag_on)$ is the marginal probability distribution of *Tag_on*. Here, the value of *Tag_off* solely depends on *Tag_on* and does not depend on any other variables. We can see a very similar relationship between *ZoneStart* and *ZoneEnd*, but *ZoneEnd* is also dependent on the choice of *Route*.

The starting node of the DAG seems to be the node *Route*. The arrows from *Route* to *ZoneStart*, *ZoneEnd*, *Direction*, and *Tag_on* show that all the later variables depend on the choice of *Route*. This makes intuitive sense as a *Route* may only be available at specific zones, direction, and time.

Other relationships from the DAG include the fact that the *ZoneStart* node directly influences the ending (*ZoneEnd*) and the direction of trips, and this choice will also affecting the choice of start time (*Tag_on*). The *Tag_on* node is only dependent on the *Direction* and *Route*, as no other variables should affect the individual decision on the time they would start their trip. Information about previous trips by the same commuter might have an influence, but as commuter identifiers are not included in the current database, there is no information available regarding this data. The shape learning

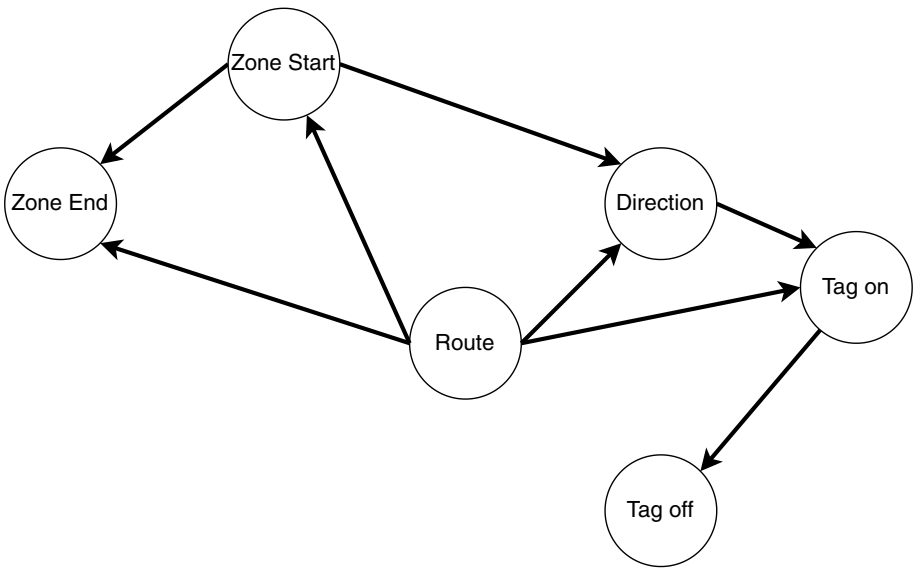


Fig. 2 Dependency structure learned by the Hill Climbing Algorithm

algorithm has thus fitted a coherent, sensible dependency structure that allows for the decision processes of commuters to be understood.

The conditional probability distributions represented by the arrows in a BN can be estimated from data using maximum-likelihood or Bayesian methods. We chose Bayesian parameter estimation as this method returns relatively similar parameters in our ten replication tests using a sample of 50,000 random trips from the Smart Card dataset. Similar tests using Maximum-Likelihood returns unstable parameters and thus are not chosen for the final analysis. Uniform priors are chosen for the Bayesian parameter estimation method, where all states are deemed equiprobable. A synthetic sample of 300,000 trips was then generated from the fitted model by forward sampling.

Finally, we implement a Chi-squared test to confirm that variables from the connected nodes in our BN are significantly correlated to each other. Table 1 shows that all the arc between source and target nodes are significantly correlated.

Generative Adversarial Network Implementation

We implemented GAN (Xu et al. 2019) using the following architecture for the generative adversarial network. The specific hyperparameters have been largely found by trial-and-error experiments.

- 1000 epochs and with *batch_size* of 100. The number of epochs decides the number of iterations that GAN will perform to optimise its parameters. *batch_size* decides the number of samples used in each step
- *embedding_dim* = 128. Size of the random sample passed to the Generator in GAN.
- Dimension of the Generator *generator_dim* = (256,256,256)
- Dimension of the Discriminator *discriminator_dim* = (256,256,256)
- Learning rate of the Generator *generator_lr*= $2e-4$, and Discriminator *discriminator_lr* = $2e-4$.
- Weight decay of the Generator's *generator_decay*= $1e-6$ and Discriminator's Adam Optimiser *discriminator_decay*= $1e-6$
- Number of Discriminator's updates for each Generator update, *discriminator_steps* = 1.

Evaluation Step 1: Probabilistic Evaluation of Synthetic Data

As detailed in Sect. "Methodology", the initial stage of our evaluation procedure for the produced synthetic data involves a probabilistic analysis of the generated distributions in comparison to the real data. We posit that the synthetic data ought to exhibit equivalent probabilistic distributions to the authentic Smart Card data.

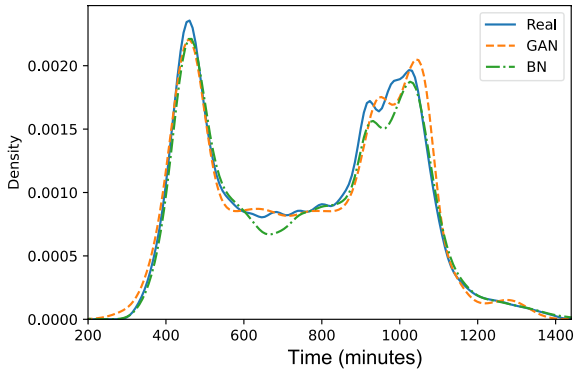
Distributions of *Tag_on* and *Tag_off*

The first distributions of interest are the distributions of *Tag_on* and *Tag_off* time, or boarding and alighting time of passengers to transit vehicles. We expect that both of these distributions are bi-modal, as there should be more people who use a transit service during morning and afternoon peak periods. Figure 3 displays the generated distributions of the tag-on times and tag-off times (in minutes from midnight) for each of the models discussed above against the real data:

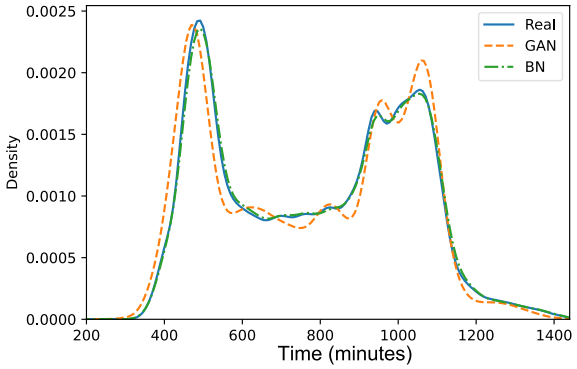
As expected, the distributions of *Tag_on* and *Tag_off* are both bi-modal with peaks during the morning and afternoon rush hours. The morning peak has a slightly higher density compared to the afternoon peak. The afternoon peak generally splits into a smaller and then the major peak. Note that *Tag_on* and *Tag_off* are both continuous variables (minutes from midnight) in the real data. While GAN learns this continuity directly from the data, we convert the variables to categorical (to a discrete 5-min interval) to facilitate learning in BN. While both BN and GAN broadly fit a mixture of normal distributions similar to the underlying data, it is clear from the plots in Fig. 3 that the distribution of the real dataset is best approximated by the BN, which has almost

Table 1 Chi-squared independence test of BN's variable nodes

Source	Target	Statistical test	p_value	Chi_square statistic	Degree of freedom
Route	Direction	True	0.0	7333.858	27
Route	Zone Start	True	0.0	1.08E+06	243
Route	Zone End	True	0.0	8.63E+05	513
Route	Tag_on	True	0.0	431652.1	6615
Direction	Zone Start	True	0.0	9.11E+03	9
Direction	Tag_on	True	0.0	9.97E+04	245
Zone Start	Zone End	True	0.0	614232.9	171
Tag_on	Tag_off	True	0.0	1.66E+07	59290



(a) Tag on time distributions for the real dataset and two models



(b) Tag off time distributions for the real dataset and two models

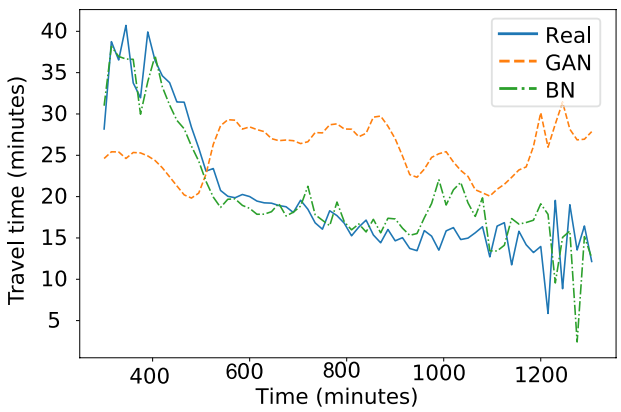
Fig. 3 Distribution of tag on/off times for the real dataset and two models

identical properties. GAN fits a similar structure, although it appears to overestimate and misplace the peaks. GAN overemphasises peaks in the data, meaning that a dataset generated from the GAN would overpredict uncommon events. Overall, we may conclude that both algorithms can fit the data well. Although BN only learns a discrete version of the data, it shows a better fit, especially at the peaks and the tails of the distribution.

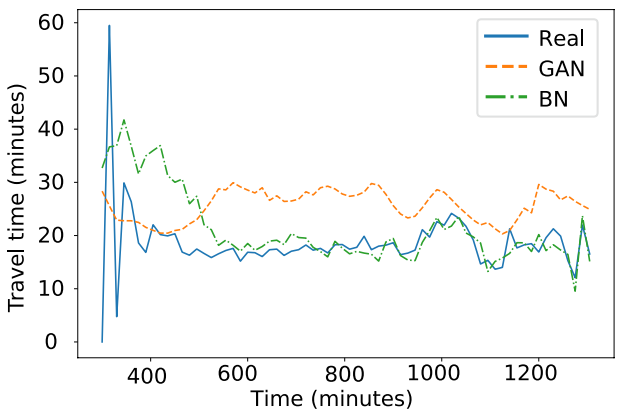
We then compare the real *Tag_on* and *Tag_off* data with generated data using a statistical significance test. The statistics will show us whether the generated data are significantly different to the real data, and which generated data (between GAN and BN) are closer to the real data. We adopt the two-sample Kolmogorov-Smirnov (KS) test to measure the distance between the cumulative density function of *Tag_on* and *Tag_off* in the generated data and the equivalent column in the real data. The output for each variable is the difference between 1 and the KS test D statistic, which means that the higher the output, the closer the generated data are to the real data. The outputs from KS test shows that BN (KS test output equals 0.987) performs better compared to GAN (KS test output equals 0.887).

Generated Travel Time

We then look at the generated travel time, versus the real data. Travel time is challenging to capture for our data synthesis models, as it is the by-product of the generated tag-on and tag-off timestamps from individual riders, as travel time has been estimated as the difference between tag-on and tag-off time. Figure 4a and b show the inbound and outbound mean travel time from BN, GAN, and the real data. The x-axis shows the time of tag on, while the y-axis shows the travel time of the associated trip. The mean travel time is calculated for each tag on interval of 15 min.



(a) Inbound travel time



(b) Outbound travel time

Fig. 4 Mean travel time at different time of day

Figure 4 shows that BN can generate very similar mean travel time to the real data. Several traffic patterns are replicated by BN, e.g., the morning peak travel time for the inbound direction and the small afternoon peak travel time for the outbound direction. Here, the real data have some noise during early morning and close to midnight, and thus, BN also captures some of the noisiness in the data. On the other hand, generated travel time from GAN seems to be around the 25 min mark, and thus, the overall trend of travel times has not been captured well by GAN. We can further confirm this by looking at the distribution of generated travel times versus the real data in Fig. 5. This figure shows that, overall, BN can replicate the travel time better than GAN, but both the algorithms struggle to capture the peak travel time. Similar to the patterns that we see in Fig. 3, GAN captures the overall distribution of travel time, but provides limited fit near the peak and tail of the distribution.

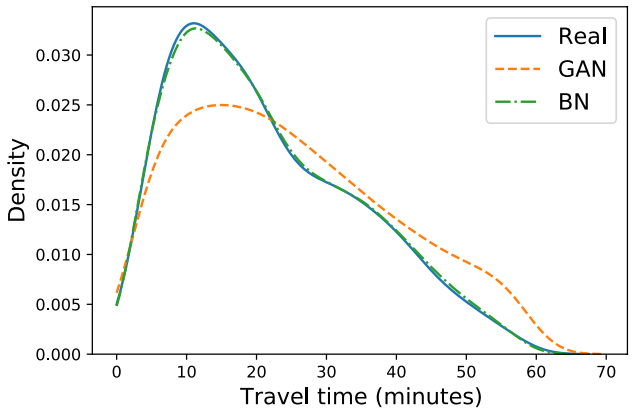


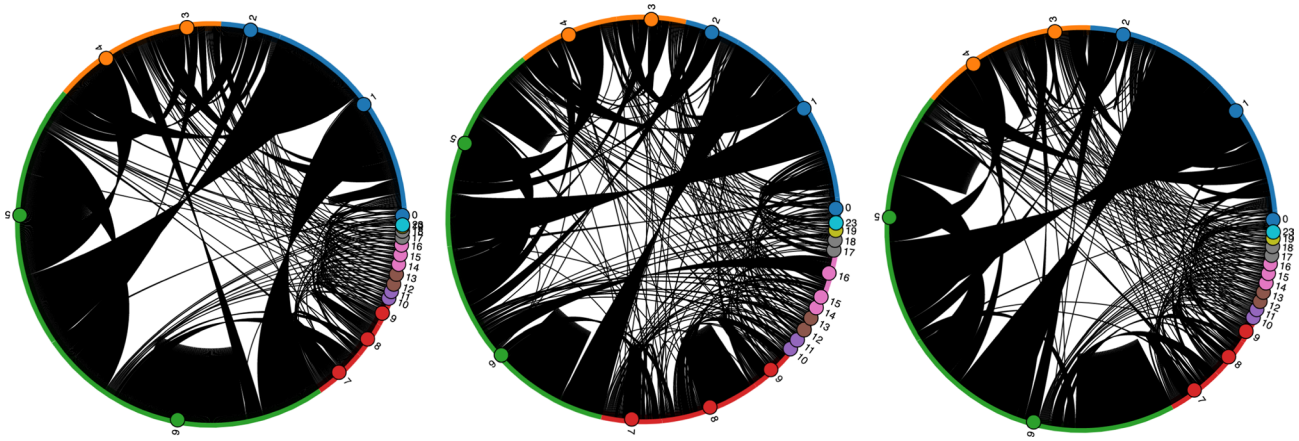
Fig. 5 Distribution of travel time

Distributions of Origin and Destination Zones

These variables are both categorical, as the zones vary from 1 (Brisbane CBD) to 23 (rural South East Queensland). If the algorithms can retain the distributions of origin–destination (OD) pairs, they can reproduce the spatial distribution of trips. Figure 6 shows three Chord diagrams of public transport trips from the real data (Fig. 6a), generated data from GAN (Fig. 6b) and generated data from BN (Fig. 6c). The chords show the number of trips between pairs public transport zones in South East Queensland, Australia. The larger the chords, the more trips are there in the data.

Both GAN and BN can replicate the overall spatial travel patterns, where the majority of the trips are between and within a few zones. Figure 6 shows that zones 1, 5, and 6 are popular zones, and there are a lot fewer trips starting or ending in zones 10–23. While both GAN and BN can replicate those patterns, BN seems to more accurately generate the proportion of trips from each zone. In the data generated by GAN, the most popular zones (zone 1, 5, and 6) are slightly less popular than in the real data set, whereas the remaining zones have a larger share than the real data.

Similar to the above section where the KS test was adopted to evaluate the fit of two continuous variables (*Tag_on* and *Tag_off*), here, we adopt another statistical test to evaluate the categorical variables *Direction*, *Route*, *ZoneStart*, and *ZoneEnd*. We implement the Chi-Squared (CS) test to compare the distributions of two discrete columns, one from the generated data and one from the real data (e.g., *Direction*, *Route*, *ZoneStart*, and *ZoneEnd*). The output for each variable is the p value from the CS tests. The value indicates the probability of the generated data having the same distribution as the real data. Thus, the higher the output, the closer the generated data are to the real data. The output of the CS test shows that the



(a) Real distribution of origin-destination pairs

(b) Generated distribution of origin-destination pairs from GAN

(c) Generated distribution of origin-destination pairs from BN

Fig. 6 Distribution of origin–destination pairs of the three models

generated categorical data from BN have a better fit compared to the data from GAN (with CS test's p value of 0.958 for BN and 0.879 for GAN, respectively).

Although the Chi-squared test is widely used for analysing discrete variables, it may not be appropriate for our dataset. This statistical test is typically based on the assumption of a normal distribution with data values exceeding five. Other distance matrices or statistical tests, such as Jaccard or Sorensen–Dice index, are not well suited for our evaluation needs, because some variables, such as origin and destination data, must be assessed jointly, and the frequency of data values is an essential factor. We then develop a new spatial distribution distance index to explore the performance of generated OD pairs and the real data, as follows:

$$d_{OD} = \sum_{i=1}^{K} |\mu_i^s - \mu_i^r|, \tag{2}$$

where μ_i^s and μ_i^r are the proportion of trips from the i_{th} pair of OD, from the synthetic and real data, respectively. μ_i^s and μ_i^r can be calculated by simply dividing the number of trips from the i_{th} pair of OD to the total number of trips in the dataset

$$\mu_i^s = \frac{\Sigma Trips_i^s}{\Sigma Trips^s} \tag{3}$$

$$\mu_i^r = \frac{\Sigma Trips_i^r}{\Sigma Trips^r}. \tag{4}$$

Unlike the popular statistical tests and distance index, the proposed spatial distribution distance index d_{OD} takes into account the patronage distributions between individual OD pair, and is free from any statistical assumptions. The generated spatial OD distribution from the synthetic data fits better with the real data if d_{OD} is smaller. The results are inline with the visual comparison on Fig. 6, as the d_{OD} value for BN is 0.14, while the d_{OD} value for GAN is 0.45.

Spatial–Temporal Travel Patterns

We then evaluate whether GAN and BN can replicate both spatial and temporal patterns in the data when they are evaluated together. Here, we look the *Origin_zone*, or the zone ID of trip origin, and *Tag_on*, or the starting timestamp of a public transport trip. Figure 7 shows the distribution of tag on timestamps on the first six travel zones in the data. It illustrates that even with one more dimension being considered (spatial zones), both GAN and BN can still replicate the temporal distribution of tag on timestamps. Both the morning and afternoon peaks are replicated in the synthetic data. Here, we can also observe that each zone has different traffic patterns that neither GAN nor BN are able to capture perfectly, e.g., the unusually high peak in the afternoon for Zone 3, or in the morning for Zone 4.

Finally, we look at the distribution of travel time at each travel zone in Fig. 8. This is the most challenging variable for GAN and BN to capture, as we are interested in a temporal by-product (travel time) that is spatially constrained

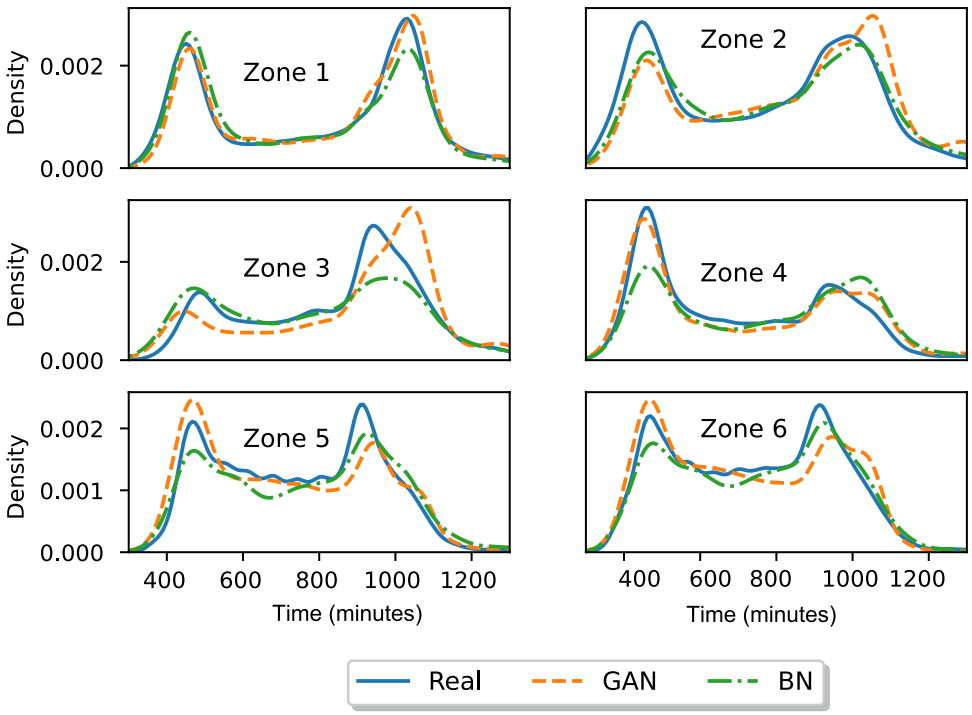


Fig. 7 Distribution of tag-on zones at different time of day

(travel zones). Figure 8 shows the real and generated distribution of travel time at the first six travel zones in the data.

Figure 8 shows that each zone has a unique distribution of travel time. Zone 1 and Zone 2 have more trips at higher travel time than the rest of the zones, while the travel time of trips from Zone 4 to Zone 6 are highly concentrated at lower values. Both GAN and BN struggle to learn the complex travel time distribution at different zones, where BN performs slightly better than GAN. The generated travel time is relatively stable across the zones. We leave the spatial learning of by-product temporal variables (e.g., travel time) to a future study, as spatial interaction data synthesis models may need to be introduced for this purpose.

Evaluation Step 2: Similarity Index for Multi-variate Synthetic Data

While we have a mixture of continuous and discrete variables in our dataset, KS and CS tests only compare individual variables, and each only works will with a single type of variable (KS test for continuous variables, and CS test for discrete variables). To adequately gauge the relative performance of the data synthesis methods discussed in this paper against each other, a similarity index was developed to compare synthetic data generated by the methods against each other and against the training data. Comparing samples from distributions over arbitrary spaces is, in general, a challenging problem. While good statistical methods such as the two-sample t test or ANOVA exist for data that is low-dimensional and normally distributed, the data here are neither normally distributed nor low-dimensional. The fact that the data are hybrid continuous-discrete data further complicates matters, as it means that defining a distance metric on the data points becomes a complex undertaking.

To this end, optimal transport theory is a proven way to measure the distance between probability distributions. For a one-dimensional space, this can be archived using the Wasserstein metric, also known as "Earth Mover's Distance", which is the distance between probability distributions. Unlike the KS or CS test, the Wasserstein metric does not require both measures to be on the same probability space. Intuitively, the Wasserstein metric is the transportation cost of shifting the probability mass of one distribution, such that it is made identical to another distribution. The shorter the distance between the distributions, the closer the distributions are to being identical. This is defined as follows:

$$W_p(\mu,\nu) := \left(\inf_{\gamma\in\Gamma(\mu,\nu)}\int_{M\times M} d(x,y)^p \,\mathrm{d}\gamma(x,y)\right)^{1/p}.$$

To estimate the distance between multivariate probability distributions, we formulate an optimal transportation problem for shifting mass from one distribution or sample to another and solving it to optimality, so distance can be found to be equal to the value of the objective function at optimality. With this in mind, we proposed a Wasserstein-based distance metric for two arbitrary dataframes of multiple variables over a defined metric space. We then solve this Linear

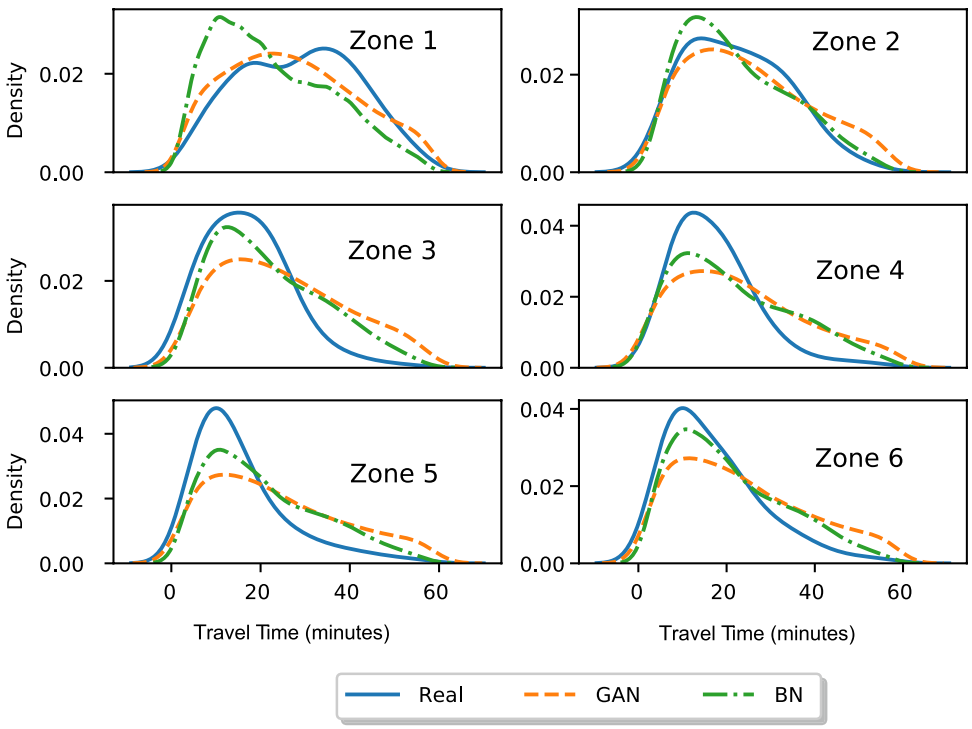


Fig. 8 Distribution of travel time at each zone

Programming problem using a Linear Programming solver (Python library PuLP and COIN-OR solver). For simplicity, we use normalised Euclidean distance to calculate the distance between data points.

Having developed an adequate way of calculating the Wasserstein metric between hybrid datasets, the next step was to estimate values for the distance between the different datasets. As calculating the values directly made comparisons difficult and was computationally prohibitive, this was performed using a bootstrapping approach: a sample of 200 data points was drawn from the training data, synthetic GAN data, and synthetic Bayesian Network data, and the distance between each pair of samples was calculated. By repeating this procedure a hundred times and taking the arithmetic means of the distances calculated between each pair of dataset samples, an estimate of the Wasserstein distances between our three datasets was calculated.

Interestingly, and in conflict with the results from the statistical method detailed above, the Wasserstein Distance implementation detailed here suggests better performance by the GAN than by BN in modelling qualities of the real data. The mean Wasserstein distance between samples from real data and the synthetic dataset created by BN is 151.52, as opposed to a distance of only 129.11 between the real data and the synthetic GAN dataset. The distance calculated between the BN data and the GAN data, 177.61, is very close to that calculated for the distance between synthetic BN data and real data. This finding is of significant interest, as it suggests that in some sense, the GAN is capturing the coupled structure of the data better than the Bayesian approach does.

A potential explanation for this finding may lie in the way in which each of the methods capture the tag-on and tag-off times in the data: in particular, the BN method can only generate data points that exist in the original datasets, whereas the GAN is capable of generating novel timestamps. This could account for the seemingly contradictory observation mentioned above: if the GAN consistently generates timestamps that are close but not identical to real timestamps, whereas the Bayesian network is forced to choose an existing timestamp to represent the data, and the observed contradiction could well be explained.

Evaluation Step 3: Privacy Validation of Synthetic Data

To demonstrate that the data generated by both methods do not violate privacy regulations by exposing real data to the end-user, samples of synthetic data were generated by the BN and GAN methods. The training data were then searched for duplicates, and this was repeated in a bootstrapping procedure.

It was found that, in a sample of 10000 synthetically generated trips using both the BN approach and the GAN approach, no real trips could be identified. Repeating this procedure ten times gave the same result, suggesting that if real trips are generated, they are exceedingly uncommon. This rate of real trips in the data thus does not pose a threat to privacy.

Discussion and Conclusion

Privacy concerns have led to the use of data anonymisation techniques to protect personal information in human activity data analysis. However, recent research has revealed the limitations of classical anonymisation techniques for big transport data, prompting the use of AI and machine learning to generate synthetic data that retain distributions and correlations without any personal records. These approaches offer a more effective solution for preserving confidentiality and data utility in research.

This paper aims to generate a synthetic version of a large transport dataset, the Smart Card data, that maintains the spatio-temporal mobility patterns of individuals while ensuring the anonymity of each data point. This objective diverges from prior research, which has primarily focused on synthetic data generation for images and texts, and has yet to adequately address transport data. The paper compares two advanced methods for synthetic data generation of human activities data, and proposes the use of Wasserstein distance to evaluate the similarity of generated synthetic data to the real data. Both methods can model the data well and generate synthetic data that are similar to the real data. BN's synthetic data are slightly better when individual variables are concerned, and even when spatio-temporal data are compared, as BN can learn the peaks in the data better. However, our proposed Wasserstein-based multivariate metric shows that when all variables are considered, GAN can produce a slightly better synthetic data.

One of the primary limitations of our approach is the coarse spatial resolution. The granularity of our synthetic data may not be sufficient for certain types of analyses, particularly those requiring geographical coordinates. Future works may include a finer-scale data synthesis for spatial data, e.g., geographically weighted neural networks Hagenauer and Helbich (2022) can be integrated into spatial-aware GAN.

Our synthetic data generation methods are also limited by the absence of travel purpose data in Smart Card data. This restricts the usability of our synthetic data for certain research inquiries, particularly those concerning travel behaviour and activity modelling. Nonetheless, it is important to emphasise that our approach is not aimed at substituting datasets that include comprehensive travel purpose

data, but rather to provide a privacy-preserving alternative for researchers who cannot access or use sensitive personal data such as Smart Card. In future studies, we could explore inferring and validating travel purpose using Smart Card data and re-evaluating the generated synthetic data to assess if it preserves travel purposes. This may require monitoring and modelling the travels of a single traveller over multiple days using unique identifiers in the dataset.

Acknowledgements The authors would like to thank Translink, the transit operator of South East Queensland, Australia for sharing the data.

Author Contributions All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Minh Kieu and Iris Brigrid Meredith. The first draft of the manuscript was written by both Minh Kieu and Iris Brigrid Meredith and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This project has received funding from the FRDF Grant at the University of Auckland, under Grant No. 3722112.

Data Availability The raw data that support the findings of this study are available from Translink, Australia, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The generated synthetic data are, however, available from the Harvard Dataverse at: https://doi.org/10.7910/DVN/9QNPFS.

Declarations

Conflict of Interest The authors would like to declare no financial and non-financial conflict of interests related to this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahmed G, Malick RAS, Akhunzada A, Zahid S, Sagri MR, Gani A (2021) An approach towards IoT-based predictive service for early detection of diseases in poultry chickens. Sustainability 13(23):13396. ISSN 2071-1050. https://doi.org/10.3390/su132313396. https://www.mdpi.com/2071-1050/13/23/13396

Axhausen KW, Gärling T (1992) Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transp Rev 12(4):323–341. ISSN 0144-1647. https://doi.org/10.1080/01441649208716826

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

Bengio Y, Thibodeau-Laufer É, Alain G, Yosinski J (2014) Deep generative stochastic networks trainable by backprop. arXiv preprint arXiv:1306.1091 [cs]

Bouman PC, Kroon LG, Schöbel A, Vervest PHM (2017) Passengers, crowding and complexity: models for passenger oriented public transport. PhD thesis, OCLC: 990177422

Briot J-P, Hadjeres G, Pachet F-D (2020) Deep learning techniques for music generation. Springer International Publishing, Computational Synthesis and Creative Systems

Choi S, Kim J, Yeo H (2021) TrajGAIL: generating urban vehicle trajectories using generative adversarial imitation learning. arXiv preprint arXiv:2007.14189 [cs, stat]

Deeva I, Andriushchenko PD, Kalyuzhnaya AV, Boukhanovsky AV (2020) Bayesian networks-based personal data synthesis. In: Proceedings of the 6th EAI international conference on smart objects and technologies for social good, pp 6–11

Drechsler J, Reiter JP (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput Stat Data Anal 55 (12):3232–3243. ISSN 0167-9473. https://doi.org/10.1016/j.csda.2011.06.006. http://www.sciencedirect.com/science/article/pii/S0167947311002076

Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv preprint arXiv:1406.2661 [cs, stat]

Guo Jessica Y, Bhat Chandra R (2007) Population synthesis for microsimulating travel behavior. Transp Res Record 2014(1):92–101

Hagenauer J, Helbich M (2022) A geographically weighted artificial neural network. Int J Geogr Inf Sci 36(2):215–235

Ismaïl S, Ahmed M, Jacques T, Bilal F, Mario C (2016) Hidden markov model-based population synthesis. Transp Res Part B Methodol 90:1–21

Jiang S, Yang Y, Gupta S, Veneziano D, Athavale S, González MC (2016) The TimeGeo modeling framework for urban mobility without travel surveys. Proc Natl Acad Sci 113 (37):E5370–E5378. ISSN 0027-8424, 1091-6490. https://doi.org/10.1073/pnas.1524261113. https://www.pnas.org/content/113/37/E5370

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. arXiv preprint arXiv:1912.04958 [cs, eess, stat]

Kieu LM, Bhaskar A, Chung E (2015) Passenger segmentation using smart card data. IEEE Trans Intell Transp Syst 16(3):1537–1548. ISSN 1524-9050. https://doi.org/10.1109/TITS.2014.2368998

Kieu LM, Bhaskar A, Cools M, Chung E (2017) An investigation of timed transfer coordination using event-based multi agent simulation. Transp Res Part C Emerg Technol 81:363–378. ISSN 0968-090X. https://doi.org/10.1016/j.trc.2017.02.018. http://www.sciencedirect.com/science/article/pii/S0968090X1730058X

Kingma DP, Welling M (2014) Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 [cs, stat]

Kuo K (2019) Generative synthesis of insurance datasets. arXiv preprint arXiv:1912.02423

Lee SG, Hickman M (2014) Trip purpose inference using automated fare collection data. Public Transp 6(1-2):1–20. ISSN 1866-749X, 1613-7159. https://doi.org/10.1007/s12469-013-0077-5

Lu R, Zhu H, Liu X, Liu JK, Shao J (2014) Toward efficient and privacy-preserving computing in big data era. IEEE Network 28(4):46–50. ISSN 1558-156X. https://doi.org/10.1109/MNET.2014.6863131

McKeown K (1992) Text generation. Cambridge University Press. ISBN 978-0-521-43802-5. Google-Books-ID: Ex6xZlxvUywC

Mendes R, Vilela JP (2017) Privacy-preserving data mining: methods, metrics, and applications. IEEE Access 5:10562–10582. ISSN 2169-3536. https://doi.org/10.1109/ACCESS.2017.2706947

Ohm P (2009) Broken promises of privacy: responding to the surprising failure of anonymization. UCLA l Rev 57:1701

Pappalardo L, Simini F (2018) Data-driven generation of spatio-temporal routines in human mobility. Data Min Knowl Discov. https://doi.org/10.1007/s10618-017-0548-4

Purdam K, Elliot M (2007) A case study of the impact of statistical disclosure control on data quality in the individual UK samples of anonymised records. Environ Plan A Econ Space 39(5):1101–1118. ISSN 0308-518X. https://doi.org/10.1068/a38335

Rasouli S, Timmermans H (2014) Activity-based models of travel demand: promises, progress and prospects. Int J Urban Sci 18(1):31–60. ISSN 1226-5934. https://doi.org/10.1080/12265934.2013.835118

Ruschendorf L (1995) Convergence of the iterative proportional fitting procedure. Ann Stat 23(4):1160–1174. ISSN 0090-5364, 2168-8966. https://doi.org/10.1214/aos/1176324703. https://projecteuclid.org/euclid.aos/1176324703

Samiul Hasan, Ukkusuri Satish V (2017) Reconstructing activity location sequences from incomplete check-in data: a semi-markov continuous-time bayesian network model. IEEE Trans Intell Transp Syst 19(3):687–698

Shrivastva KMP, Rizvi MA, Singh S (2014) Big data privacy based on differential privacy a hope for big data. In: 2014 international conference on computational intelligence and communication networks, p 776–781.https://doi.org/10.1109/CICN.2014.167

Sun L, Erath A (2015) A Bayesian network approach for population synthesis. Transp Res Part C Emerg Technol 61:49–62. ISSN 0968-090X. https://doi.org/10.1016/j.trc.2015.10.010. https://www.sciencedirect.com/science/article/pii/S0968090X15003599

Sweeney L (2001) Computational disclosure control: a primer on data privacy protection. Thesis, Massachusetts Institute of Technology. https://dspace.mit.edu/handle/1721.1/8589

Translink (2019) South Eastern busway network map. Technical report

Wallner G, Kriglstein S, Chung E, Syeed K (2018) Visualisation of trip chaining behaviour and mode choice using household travel survey data. Public Transp. https://doi.org/10.1007/s12469-018-0183-5

Willenborg L, de Waal T (2001) Elements of statistical disclosure control. Lecture Notes in Statistics. Springer-Verlag, New York. ISBN 978-0-387-95121-8. https://doi.org/10.1007/978-1-4613-0121-9. https://www.springer.com/gp/book/9780387951218

Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K (2019) Modeling tabular data using conditional GAN. Adv Neural Inf Process Syst 32:7335–7345. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html

Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP (2020) Generation and evaluation of privacy preserving synthetic health data. Neurocomputing. ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2019.12.136. http://www.sciencedirect.com/science/article/pii/S0925231220305117

Yan-Feng L, Jinhua M, Liu Y, Yuan-Jian Y, Hong-Zhong H (2015) Dynamic fault tree analysis based on continuous-time bayesian networks under fuzzy numbers. Proc Inst Mech Eng Part O J Risk Reliab 229(6):530–541

Yin M, Sheehan M, Feygin S, Paiement J-F, Pozdnoukhov A (2017) A generative model of urban activities from cellular data. IEEE Trans Intell Transp Syst 19(6):1682–1696

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.