Synthetic generation of individual transport data: the case of Smart Card data

Minh Kieu, Iris Brighid Meredith, Andrea Raith
University of Auckland, New Zealand

March 11, 2022

1 Introduction

The acquisition and dissemination of individual data are key for research in many disciplines, including social simulation. 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. However, governments across the globe have increasingly become concerned over privacy and the exchange of personal data. 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). This poses a major risk of privacy breaches for vulnerable individuals and thus prevents the wider dissemination of personal data to the wider research community and limits the impacts of research on policymaking. On the other hand, as the risks of personal data disclosure increase, the alterations made by data owners may impact the usefulness of the released data.

To address the limitations of standard de-identification measures, literature has offered various approaches aiming at generating partially synthetic or fully synthetic data from real data. The idea is to retain the probability distributions in the data, but each synthetic data sample does not represent a real person in the raw data. Synthetic data enables 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 policymakers, 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), synthetic data is often poorly understood in transportation. Human mobility-related data in transport is relatively unique compared to popular personal data such as census data, health records or financial data because individual transport data such as Smart Card data often have operational information such as travel routes and modes, which are strictly spatially constrained, e.g. train travels occur only to and from train stations. Analysts generating and working with synthetic transport data must be aware of the confidentiality of this spatial element while aiming to retain the spatial information in the data.

This paper compares two of the most advanced methods for data modelling and synthetic data generation: Bayesian Network and Generative Adversarial Network for the generation of the most popular individual data in transportation: the Smart Card data. Smart Cards have become the de facto standard for modern public transport systems. The availability of Smart Card data has recently enabled novel research in intelligent transport systems, such as the analysis of travel behaviours (Kieu et al. 2015, 2018), inference of trip purposes (Lee and Hickman 2014), or intention to transfer (Kieu et al. 2017). However, the research community has not widely benefited from the ubiquitous availability of Smart Card data to support decision making while policymakers, who may have access to the raw data, have not yet been informed by the cutting-edge research on their data. The framework in this paper connects Smart Card

data owners to a much wider community of researchers through synthetic data modelling and generation. It enables researchers to work on a synthetic dataset that is reasonably similar to the real data, having the same distributions and retaining the spatial-temporal activity sequence in the real data, but with data points not representing real people. On the other hand, this paper provides public transport agencies, research centres, local councils and other Smart Card data owners with a better alternative for public data dissemination. The scientific contributions of this paper are three-fold:

- We introduce a new data pipeline to process raw Smart Card data into sequential spatiotemporally constrained trip data
- We apply a Generative Adversarial Network, 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

2 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 zero-sum game (based on game theory), 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).

Among the latest GAN-based algorithms in the literature, we adopt Tabular Conditional GAN (CTGAN) for modelling and generating of Smart Card data (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 insurance data (Kuo 2019).

3 Bayesian Networks

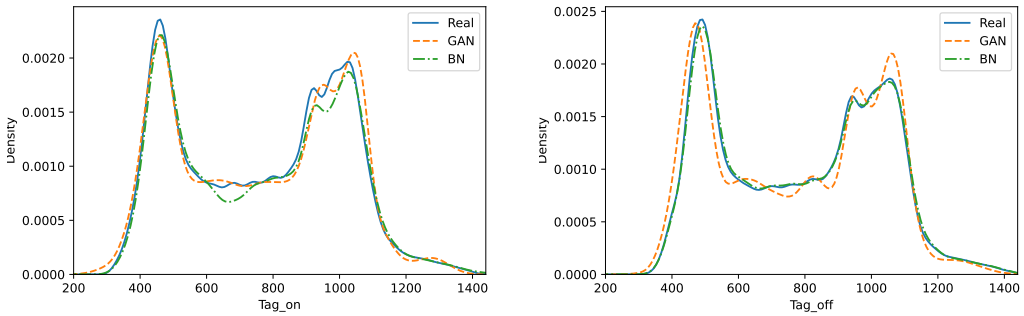
A Bayesian Network is a directed acyclic graph with a variable and a conditional probability associated with each node: the distribution for each variable is conditioned on the variables upstream of it in the DAG. A joint probability distribution for the variables can then be fitted on the graph and sampled from, generating a synthetic sample. In this paper, shape learning on the dataset was performed using the hill-climbing and constraint-based search approaches, with parameter fitting being performed by expectation-maximisation and forward sampling being used to generate the synthetic datasets.

4 Evaluation of generated data

In this section, we analyse and compares the generated Smart Card data from BN and GAN. We hypothesise that the generated data should have the same probabilistic distributions as the real Smart Card data.

Figure 1 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:

Both BN and GAN broadly fit a mixture of normal distributions similar to the underlying data, it is clear from the plots that the real dataset's distribution is best approximated by the BN, which has almost



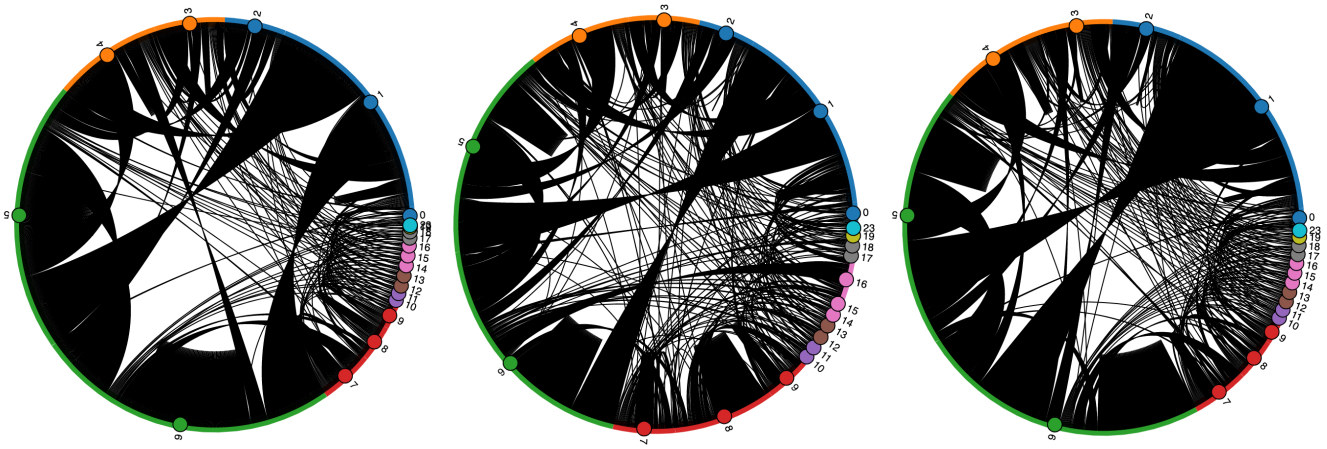
(a) Tag on time distributions for the real dataset and three models

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

Figure 1: Tag off time distributions for the real dataset and three models

identical properties. GAN fits a similar structure, though it appears to overestimate and misplaces the peaks. GAN overemphasise peaks in the data, meaning that a dataset generated from the GAN would underpredict uncommon events.

We then look at the distribution of origin and destination zones. These variables are categorical, as the zones vary from 1 to 23. If the algorithms can retain the distributions of origin distribution, they can reproduce the spatial distribution of trips. Figure 2 shows three Chord diagrams of public transport trips from the real data (Figure 2(a)), generated data from GAN (Figure 2(b)) and generated data from BN (Figure 2(c)). The larger the chords, the more trips are there in the data.



(a) Real distribution of Origin-distribution

(b) Generated distribution of Origin-distribution from GAN

(c) Generated distribution of Origin-distribution from BN

Figure 2: Tag off time distributions for the real dataset and three models

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 2 show that zone 1, 5 and 6 are popular zones, and there are a lot fewer trips started or ended in zones 10 to 23. While both GAN and BN can replicate those patterns, BN seems to more accurately generate the proportion of trips from each zone. In GAN the most popular zones (zone 1,5 and 6) are slightly less popular, whereas the remaining zones have a larger share than the real data.

Finally, we look at the distribution of travel time at each travel zone in Figure 3. 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 (travel zones). Figure 3 shows the real and generated distribution of

travel time at the first 6 travel zones in the data.

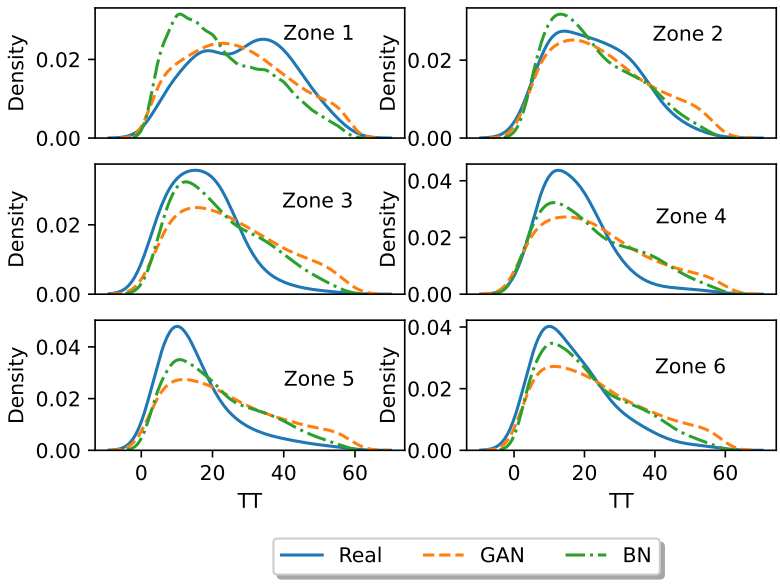


Figure 3: Distribution of travel time at each zone

Figure 3 shows that each zone has a unique distribution of travel time. Zone 1 and Zone 2 has more trips at higher travel time than the rest, while the travel time of trips from Zone 4 to 6 is highly concentrated at lower values. Both GAN and BN struggle to learn the complex travel time distribution at different zones, with BN performing slightly better than GAN. The generated travel time is relatively stable across the zones. We leave the spatial learning of by-product temporal variable (e.g. travel time) to a future study, where spatial interaction data synthesis models may need to be introduced for this purpose.

5 Conclusion and future works

This abstract describes the current progress of an ongoing project "Synthetic Big Data of Human Activities (SynAc)". The comparison between Bayesian Network (BN) and Generative Adversarial Network (GAN) shows that both methods can model and generate data that have the same distributions with the real data, both spatially and temporally. The synthetic data from Smart Card can be used as the synthetic population for an Agent-Based Models of public transport.

The next step in SynAc is to retain the sequential structure of transportation data, such as individual travels at a certain time from one area to another. This structure expresses the individuality of each person in as much as their activities are associated with travelling. The sequential travel activity from Smart Card data is even more challenging to synthesise as a person's travel itinerary will be incomplete if some of the travel is not done on public transport. We are currently exploring BN, GAN and various other methods in synthetic data modelling and generation for sequential transport data.

References

Jean-Pierre Briot, Gaëtan Hadjeres, and François-David Pachet. *Deep Learning Techniques for Music Generation*. Computational Synthesis and Creative Systems. Springer International Publishing, 2020. ISBN 978-3-319-70162-2. doi: 10.1007/978-3-319-70163-9. URL `https://www.springer.com/gp/book/9783319701622`.

Jörg Drechsler and Jerome P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243, December 2011. ISSN 0167-9473. doi: 10.1016/j.csda.2011.06.006. URL http://www.sciencedirect.com/science/article/pii/S0167947311002076.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv: 1406.2661.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]*, March 2020. URL http://arxiv.org/abs/1912.04958. arXiv: 1912.04958.

L. M. Kieu, A. Bhaskar, and E. Chung. Passenger Segmentation Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1537–1548, June 2015. ISSN 1524-9050. doi: 10.1109/TITS.2014.2368998.

Le Minh Kieu, Ashish Bhaskar, Mario Cools, and Edward Chung. An investigation of timed transfer coordination using event-based multi agent simulation. *Transportation Research Part C: Emerging Technologies*, 81:363–378, August 2017. ISSN 0968-090X. doi: 10.1016/j.trc.2017.02.018. URL http://www.sciencedirect.com/science/article/pii/S0968090X1730058X.

Le Minh Kieu, Yuming Ou, and Chen Cai. Large-scale transit market segmentation with spatial-behavioural features. *Transportation Research Part C: Emerging Technologies*, 90:97–113, May 2018. ISSN 0968-090X. doi: 10.1016/j.trc.2018.03.003. URL http://www.sciencedirect.com/science/article/pii/S0968090X1830278X.

Kevin Kuo. Generative synthesis of insurance datasets. *arXiv preprint arXiv:1912.02423*, 2019.

Sang Gu Lee and Mark Hickman. Trip purpose inference using automated fare collection data. *Public Transport*, 6(1-2):1–20, April 2014. ISSN 1866-749X, 1613-7159. doi: 10.1007/s12469-013-0077-5. URL http://link.springer.com/10.1007/s12469-013-0077-5.

Kathleen McKeown. *Text Generation*. Cambridge University Press, June 1992. ISBN 978-0-521-43802-5. Google-Books-ID: Ex6xZlxvUywC.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems*, 32:7335–7345, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html.