Enhancing MaaS Personalisation Through Synthetic Data Generated from a Tabular Large-Scale Mobility Dataset

Francesco Maria Turnoa[1(✉)], Irina Yatskiva[1], Maksim Ilina[1], Sigita Lapinaa[1], and Luca Gillib[2]

[1] Transport and Telecommunication Institute, Riga, Latvia
turno.fm@tsi.lv
[2] Clearbox AI, 10129 Turin, Italy

Abstract. In recent years, the concept of Mobility-as-a-Service has significantly impacted the transportation sector by integrating diverse modes of transport into a user-friendly experience. The advancement of human mobility patterns has been facilitated by the utilisation of mobile sensing technologies, but this progress has also raised concerns regarding privacy and the management of data. This study suggests increasing the applicability of human mobility data by generating synthetic data with deep learning models trained on the existing dataset. Our approach aims to enhance the practicality of human mobility data. The produced synthetic data encompasses real-world dynamics and give possibility to develop and evaluate the algorithms for personalised travel recommendations, while safeguarding sensitive information. Exploring this domain has the potential to bring about a paradigm shift in the field of mobility solutions that prioritise privacy, efficiency, and user satisfaction, ultimately leading to the development of a sustainable urban mobility framework.

Keywords: Human Mobility Data · Personalisation · Generative Artificial Intelligence · Utility · Fidelity

1 Introduction

The concept of Mobility-as-a-Service (MaaS) has gained significant attention in recent years, revolutionising the way people move by integrating various modes of transportation into a seamless user experience. The mobile sensing technologies continuously collect a wealth of data about people's movements, interactions, and behaviours, resulting in a massive influx of data. However, while this abundance of data has enormous potential for improving our understanding of human mobility patterns, it also poses privacy, data security, and data management challenges.

The introduction of *synthetic data generation* (SDG) techniques opens a wide range of transformative possibilities in the field of mobility research, especially when it comes to delivering personalised travel recommendations. *Synthetic data* (SD) can be defined as artificially generated data that uses pre-existing datasets or models to accurately

C. McNally et al. (Eds.): TRAconference 2024, LNMOB, pp. 565–571, 2026.
https://doi.org/10.1007/978-3-031-88974-5_82

emulate the statistical attributes and characteristics present in real-world data [1]. This is especially useful in situations where obtaining actual data would be prohibitively expensive, time-consuming or impossible [2]. Its importance originates from the intrinsic adaptability in addressing a set of problems such as data augmentation [3], imputation of missing data [4], restoring "fairness" in biased data [5], and ensuring confidentiality [6]. Furthermore, it enables researchers to develop and test algorithms and applications in a controlled environment, removing the need to expose sensitive data.

Current advances in SDG are becoming popular in various domains including ecology, computer vision, industrial engineering etc. However, it isn't yet extensively explored in the mobility area. In this study, we propose a approach to increase the applicability of human mobility data by generating synthetic data with deep learning (DL) models trained on the existing dataset. Next Section describes applied research methods and Sect. 3 is devoted to results' presentation. The final goal of this research is to analyse the place of SDG techniques within the realm of personalised travel recommendation algorithms. By intertwining real data from actual trips with synthetic data, the research creates a dynamic and adaptable framework for future travel suggestions.

2 Data and Methods

Data Description. The rapidly growing field of human mobility data (HMD) benefits from the Sussex-Huawei Locomotion (SHL) dataset, which captures the human movement in various contexts and environments [7]. The latter explores human locomotion from running and walking to biking and public transport. The data collection process involved three people and reflected the spontaneity and unpredictability of human mobility through everyday activities like long-distance travel and museum visits. For complete movement data capture, Huawei Mate 9 smartphones were placed at the hands, torso, hips, and bags, with the purpose of measuring pressure, ambient light, acceleration, and GPS coordinates. The dataset's annotations, which describe movement modes, environmental and situational contexts like road conditions, traffic scenarios, and social interactions, demonstrate its high variability.

Personalising services means adapting to an individual's behaviour. Therefore, focusing on a single user's (*User 1*) GPS locations and activity labels is crucial in this context. Due to its narrow focus, the dataset contains every move, turn, and pause, as well their unique behaviours, routines, and social habits. The study's 391 h over five months show its focus on User 1's mobility's complexities and patterns. This longitudinal method considers seasonal fluctuations, changing habits, and external mobility influences, giving a more complete picture than short-term datasets. It also allows a more in-depth analysis without the noise and disparities of a multi-user dataset by removing the broader variability introduced by multiple users.

SDG Model Architecture. A notable development in the field of unsupervised machine learning is the introduction of *generative adversarial networks* (GANs) [8]. GANs consist of two neural components: the *discriminator*, which verifies the authenticity of the SD by comparing it to real data, and the *generator*, which generates SD. The training process continuously improves both the discriminator and the generator. A model capable in tabular data synthesis is of course necessary when it comes to the SHL dataset,

which integrates location and activity labels. On the other hand, the *conditional tabular generative adversarial network* (CTGAN), is a model specifically designed to handle categorical attributes found in tabular data and ensuring consistent and stable learning [9, 10], which makes this model an ideal candidate for generating data that closely resembles the characteristics of our mobility dataset (Table 1).

Table 1. Data structure of input data for SDG task

Attribute	Timestamp	Latitude	Longitude	Activity label
Type	Datetime format (YYYY-MM-DD hh:mm:ss)	Float	float	String
Example	2017-03-01 13:58:18	50.86075	−0.09689	Still_Sit_Inside

The model is initialised with a *random noise vector* (sampling from a *standard Gaussian distribution*) that is the primary input for the generator. The vector is propagated over multiple *fully connected* layers. The *Rectified Linear Unit* (ReLU) is the *activation function* that is used in each of these layers. It is essential for introducing the non-linearities that the generator needs to properly represent complex data distributions. Each layer has *Batch Normalisation* (BN) applied to it to improve training stability and convergence. BN standardises intermediate activations and reduces the problem of internal *covariate shift*. To accommodate the intrinsic heterogeneity of tabular data, the generator's terminal layer uses a bifurcated activation strategy. For discrete attributes, such as "activity label", the *Gumbel-Softmax* activation is employed because it facilitates gradient-based methods that allow for optimisation in discrete spaces. The *hyperbolic tangent* (tanh) function is utilised for continuous attributes as "latitude", which precisely mirror the distributions found in actual data.

By using the same amount of densely connected layers, the discriminator is made to switch to its adversarial counterpart. Each layer in the model makes use of the *LeakyReLU* activation function, which is an altered form of the standard ReLU. Including this function allows for the possibility of a slight negative gradient during the times when the unit is not in use. This choice solves the problem of "diminishing neuron" activation, ensuring that gradients flow consistently during the *backpropagation* process. These layers also include *dropout* mechanisms that, with each training cycle, probabilistically deactivate a subset of neurons.

Ultimately, to guarantee *Lipschitz continuity*, CTGAN uses the *Wasserstein loss function with gradient penalty*, which enables the synthesised data to statistically match the original dataset and adhere to the intricate spatial "coordinates-activity" dynamics present in actual activity recordings.

To evaluate the *coherence* of suggested approach was used set of metrics proposed in [11] and investigated the quality of SD generation on HMD. First metric is ***Absolute Semantics***, which refers to the understanding of the meaning of each location in a trajectory. This can be evaluated by analysing the locations and types of activities

present in both types of data and by comparing their distributions to assess representativeness. ***Marginal Distributions*** evaluates the distribution of individual variables in the SD compared to the original dataset. If these distributions are similar, it suggests that the SD maintains the statistical properties of the original data. The ***Maximum Mean Discrepancy*** (MMD) measures the difference between distributions of SD and the original. ***Jensen-Shannon Divergence*** (JSD) quantifies the similarity between two probability distributions and is a symmetrized and smoothed version of the *Kullback-Leibler divergence.* A smaller MMD and JSD mean the distributions are more similar. ***Relative Semantics*** entails understanding a location's meaning relative to other locations in a trajectory, involving transitions analysis between different types of locations and evaluating their logic. *Pairwise semantic distances*, involving distances calculations between location pairs based on their semantic meanings, can measure this. By ensuring that SD closely mirrors real data in terms of semantics and spatial-temporal patterns, ensure that personalised mobility algorithms are built on robust and meaningful data.

3 Results

The following results demonstrate the DL model's ability to generate SD.

Absolute Semantics: the scatterplots on Fig. 1 depict the spatial distribution of activities in datasets. The original dataset exhibits distinct clusters of activities, suggesting specific geographical zones where certain activities predominantly occur. For example, certain latitudes and longitudes have a high concentration of "Bike" activities, which could imply popular biking routes or parks. Other clusters might indicate residential or commercial areas based on the nature of the activities. The generated by CTGAN SD aims to replicate these patterns. While it does capture the general spatial structure and distribution of activities, there are subtle differences. Some activity clusters in the SD are more dispersed than in real, while others are denser. This divergence can be attributed to the inherent randomness and generative nature of SD production. However, the overall spatial semantics are largely preserved in the SD, making it a valuable resource for analysis that doesn't compromise individual data points from the original set.

Marginal Distributions: the Kernel Density Estimation plots for marginal latitude and longitude distributions show distinct patterns that demonstrate the SD generation process's accuracy (Fig. 2). Blue curves represent the original data and show peaks and distributions of key geographical areas where activities are concentrated. The SD's red curves closely match these patterns, indicating that the SD generation captured the original dataset's spatial nuances. Although some latitude and longitude ranges deviate, the overall congruence is good, and the distributions are similar.

MMD and JSD: values are presented in Table 2. "Latitude" has a very low MMD, indicating that the original and SD sets have similar means. JSD value of 0.35, which is not negligible, but not excessive, supports this. It means that the probability distributions for this attribute are similar. The MMD value for "Longitude" is slightly higher, but still small, indicating close means and the JSD value of 0.42 suggests that the distributions differ, but the SD still closely matches the original data. For the "Activity label" MMD and JSD are extremely low. Hence, the distributions of this categorical variable in the

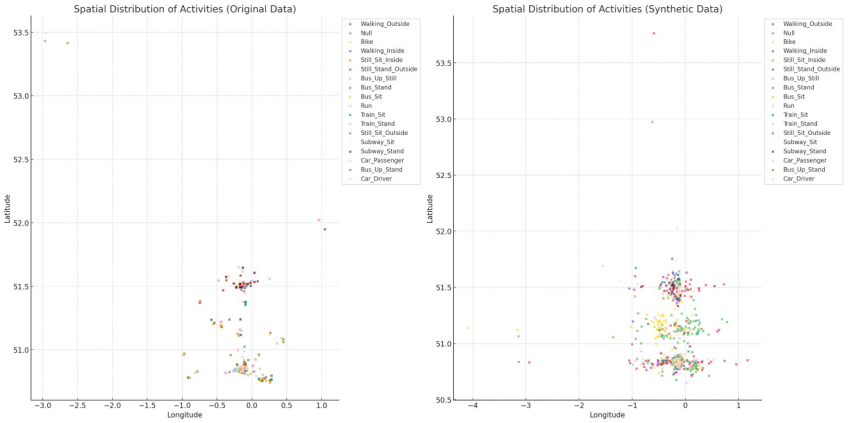

Fig. 1. Spatial distribution of activities: real user (left), simulated agent (right). Created by Authors in Python.

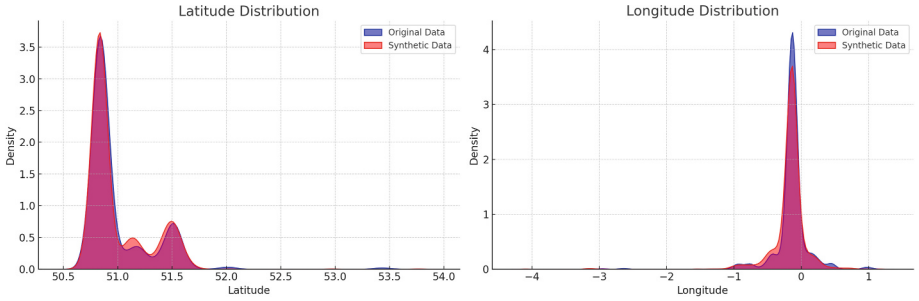

Fig. 2. Kernel Density Estimation plots for Latitude and Longitude distributions. Created by Authors in Python.

original and SD sets are very similar, which means that CTGAN replicated well the activity label distribution from the original dataset.

Table 2. Summary of MMD and JSD for all attributes.

Attribute	MMD	JSD
Latitude	0.000077	0.35
Longitude	0.00052	0.42
Activity label	0.00072	0.02

Relative Semantics: the matrices (Fig. 3) capture the transition probabilities between different activity labels. In the original data heatmap, distinct patterns of activity transitions can be discerned, with certain activity combinations exhibiting higher probabilities,

as indicated by the deeper shades of blue. These patterns provide insights into the inherent structure and sequence in which activities tend to occur. The SD heatmap seeks to emulate these patterns. Most of the transition patterns in the original data are mirrored in the SD, as evident from the consistency in colour shades. This suggests that the CTGAN has effectively captured the dynamics of activity transitions, preserving the intricate relationships and sequences between activities.

4 Conclusion

There results outlined in this paper highlight the capabilities of CTGAN, in generating synthetic human mobility data of exceptional quality. These datasets accurately reflect the inherent characteristics and dynamic patterns of mobility, providing a solid basis for future research. Within this framework, the reliability of the generative task was assessed by comparing the artificial dataset with the real dataset using statistics-based metrics. This serves as a pivotal reference point for evaluating the accuracy and efficacy of GAN-based models. Hence, this model has potential applications in the field of Mobility as a Service (MaaS), which include the ability to enhance route planning, analyse individual mobility patterns, identify preferred modes of transportation, and simulate user behaviour in various transportation scenarios.

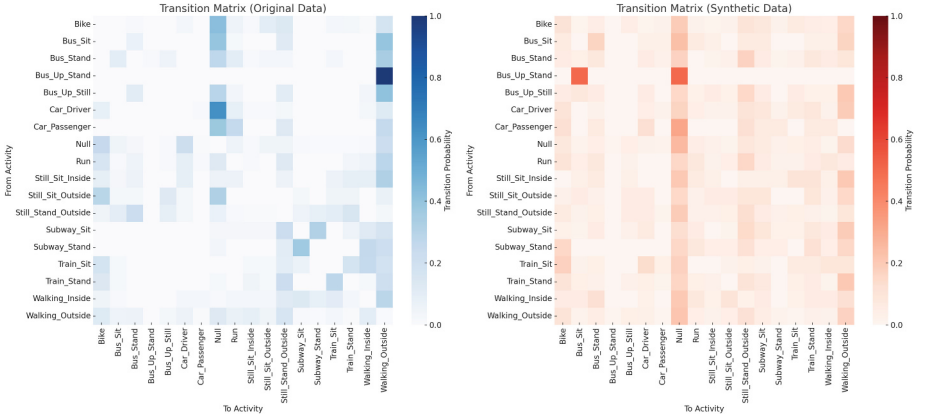

Fig. 3. Heatmaps highlighting activity changes for real and synthetic dataset. Created by Authors in Python.

Acknowledgements. The research in the frame of project DCODE has received funding from the EU Horizon 2020 RI programme under the Marie Skłodowska–Curie grant agreement No 955990.

References

1. Dankar, F.K., Ibrahim, M.: Fake it till you make it: guidelines for effective synthetic data generation. Appl. Sci. **11**(5), 2158 (2021)
2. Nikolenko, S.I.: Synthetic data for deep learning, SOIA (174), Springer (2021)
3. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using GAN for improved liver lesion classification. In: Proceeding of 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 289–293 (2018)
4. Chen, Y., Lv, Y., Wang, F.: Traffic flow imputation using parallel data and generative adversarial networks. IEEE Trans. Intell. Transp. Syst. **21**(4), 1624–1630 (2019)
5. Xu, D., Yuan, S., Zhang, L., Wu, X.: FairGAN: fairness-aware generative adversarial networks. In: Proceeding of 2018 IEEE International Conference on Big Data, pp. 570–575 (2018)
6. Dash, S., Dutta, R., Guyon, I., Pavao, A., Bennett, K.P.: Privacy preserving synthetic health data. In: ESANN 2019 (2019)
7. Gjoreski, H., et al.: The university of Sussex Huawei locomotion and transportation dataset for multimodal analytics with mobile devices. IEEE Access **6**, 42592–42604 (2018)
8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
9. Xu, L, Veeramachaneni, K.: Synthesising tabular data using generative adversarial networks. arXiv preprint arXiv:1811.11264 (2018)
10. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modelling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
11. Eigenschink, P., Reutterer, T., Vamosi, S., Vamosi, R., Sun, C., Kalcher, K.: Deep generative models for synthetic sequential data: a survey. IEEE Access (2023)