Collaborative generative adversarial networks for fusing household travel survey and smart card data to generate heterogeneous activity schedules

Huichang Lee[1], Prateek Bansal[2,3], Khoa D. Vo[2,4], Eui-Jin Kim[5,*]

[1] Department of Civil and Environmental Engineering, Seoul National University, Republic of Korea
[2] Singapore-ETH Centre, Future Cities Lab Global Programme, Singapore Hub, Singapore
[3] Department of Civil and Environmental Engineering, National University of Singapore, Singapore
[4] Faculty of Information Technology, Industrial University of Ho Chi Minh City, Vietnam
[5] Department of Transportation System Engineering, Ajou University, Republic of Korea
*Corresponding author

Extended abstract submitted for presentation at the 12th Triennial Symposium on Transportation Analysis conference (TRISTAN XII)
June 22-27, 2025, Okinawa, Japan

February 28, 2025

Keywords: Deep generative model, Activity-based model, Data fusion, Smart card

1 INTRODUCTION

Traditional activity-based models (ABMs) relying on household travel survey (HTS) data suffers from low sampling rates due to high survey costs and this leads to the "zero-cell problem", the exclusion of activity schedules that exist in the true population. This issue results in the generation of less diverse activity schedules, referred to as the problem of "low spatiotemporal heterogeneity", which leads to inaccurate demand forecasts that overlook sparse mobility patterns. Additionally, the long data collection cycle of HTS makes it difficult to capture the up-to-date changes.

Deep generative models (DGMs), such as variational autoencoder (VAE) and generative adversarial networks (GAN), are well-suited for modeling high-dimensional joint distributions. Previous research has demonstrated that DGMs achieve greater heterogeneity in the population synthesis using HTS data (Borysov *et al.*, 2019). However, the HTS data used for training DGMs is lacking in heterogeneity and outdated, limiting their performance in activity scheduling.

This study proposes a novel DGM-based data fusion method for generating activity schedules, which preserves the comprehensive information of HTS while maintaining the high spatiotemporal heterogeneity and up-to-date nature of SC data. The proposed method, collaborative generative adversarial networks (CollaGAN), incorporates two discriminators during the data fusion process, ensuring that the strengths of both datasets are preserved. To enhance the feasibility of the generated activity schedules, three novel loss functions are designed. This approach produces spatiotemporally heterogeneous and up-to-date activity schedules.

2 METHODOLOGY

VAE compresses high-dimensional data into a lower-dimensional latent space, such as a multivariate normal distribution, to model its underlying distribution (Kingma, 2013). Based on this VAE framework, the spatiotemporal attributes of HTS and SC data are mapped into the mean and variance of the latent distribution $p(z)$ through a single encoder. The encoder forces the outputs to follow

$p(z)$ by minimizing the following Kullback-Leibler (KL) divergence loss, which measures the distance between two probabilistic distributions:

$$\mathcal{L}_{KL,\text{hts}} = -D_{KL}[q_\phi(z|T_{\text{hts}}, S_{\text{hts}})||p(z)] \quad (1)$$

$$\mathcal{L}_{KL,\text{sc}} = -D_{KL}[q_\phi(z|T_{\text{sc}}, S_{\text{sc}})||p(z)] \quad (2)$$

Here, T and S represent the temporal attributes (i.e., start time and duration) and the spatial attributes (i.e., location) of activity schedules, respectively. Ultimately, the distribution learned by the encoder becomes $q_\phi(z|T,S)_{\text{fus}}$, a harmonized form of $q_\phi(z|T_{\text{hts}}, S_{\text{hts}})$ and $q_\phi(z|T_{\text{sc}}, S_{\text{sc}})$. We design a harmony loss function to ensure that both datasets maintain consistency within the uniform latent space, allowing efficient information transfer between HTS and SC:

$$\mathcal{L}_{KL,har} = -D_{KL}[q_\phi(z|T_{\text{sc}}, S_{\text{sc}})||q_\phi(z|T_{\text{hts}}, S_{\text{hts}})] \quad (3)$$

During the inference process, the generator generates the attributes of activity schedules from the latent vector sampled from the fused distribution. In this process, we train the generator in a semi-supervised manner to generate the qualitative attributes. The generator restores travel mode M and activity purpose A from HTS data and infers M and A from SC data. It is trained to minimize the following reconstruction loss between the inputs from both data sources and the generated data:

$$\mathcal{L}_{recon,\text{hts}} = \left\|T_{\text{hts}} - \hat{T}_{\text{hts}}\right\| + \left\|S_{\text{hts}} - \hat{S}_{\text{hts}}\right\| + \left\|M_{\text{hts}} - \hat{M}_{\text{hts}}\right\| + \left\|A_{\text{hts}} - \hat{A}_{\text{hts}}\right\| \quad (4)$$

$$\mathcal{L}_{recon,\text{sc}} = \left\|T_{\text{sc}} - \hat{T}_{\text{sc}}\right\| + \left\|S_{\text{sc}} - \hat{S}_{\text{sc}}\right\| \quad (5)$$

$\|\cdot\|$ represents the categorical cross-entropy. The generator can reconstruct the original data from the latent space by minimizing the reconstruction loss. When $z \sim p(z)$ is fed into the generator, it generates activity schedules that conform to the distribution $p(T,S,M,A)_{\text{fus}}$.

We involve discriminators D_{hts} and D_{sc} from each dataset's perspective in an adversarial game with the generator G to ensure that the comprehensiveness of HTS data and the heterogeneity of SC data are preserved during the data fusion process. Based on the strengths of each data, D_{hts} evaluates the set of T, M, and A, whereas D_{sc} concentrates on the combination of T and S. The generator and the two discriminators form a CollaGAN structure and engage in a min-max game with the following:

$$\min_G \max_{D_{hts}, D_{sc}} V(D,G) = E\left[log\, D_{hts}\left(\psi_{hts} G(z)\right)\right] + E[log(1 - D_{hts}(\psi_{hts} G(z)))] + E\left[log\, D_{sc}\left(\psi_{sc} G(z)\right)\right] + E[log(1 - D_{sc}(\psi_{sc} G(z)))] \quad (6)$$

where ψ_{hts} and ψ_{sc} are functions that extract the attributes to be assessed by D_{hts} and D_{sc}, respectively. Through this process, we ensure that the strengths of each dataset are retained in the fused joint distribution learned by the generator.

To prevent the generation of infeasible attribute combinations that are unlikely to exist, we apply regularizations during the training process. Boundary loss (Kim and Bansal, 2023) in Equation 7 measures the distance between the training samples and the generated samples at the boundary of the sample space, distinguishing infeasible samples. Let $\hat{X}_j$ represent m generated data points in a mini-batch, while X refers to the entire training dataset with a size of N.

$$R_{BD}(\hat{X}, X) = \frac{1}{m}\sum_{j=1}^{m} \min_{i \in \{1:N\}} \left(\text{Dist}\left(\hat{X}_j, X_i\right)\right) \quad (7)$$

We implement a regularization loss that acts as the expert-designed constraints found in traditional econometric models for activity scheduling. We define the range of feasible attribute combinations and apply a significant penalty whenever the generator produces samples outside that range, ensuring the generated activity schedules align with domain knowledge. For example, we assume every first trip to be a home-based trip and apply the loss if the first activity purpose is home

return. Since SC data only contains transit records for transit users, the fused joint distribution can become biased toward transit users and shorter trip chain lengths. To correct this bias, we apply rejection sampling using the fused joint distribution estimated by the generator as the proposal distribution.

3 RESULTS AND DISCUSSION

We apply the feasibility and heterogeneity metrics, proposed by Kim and Bansal (2023). Feasibility (Equation 8) measures how well the generated data mimics the population. Heterogeneity (Equation 9) indicates the degree to which the generated data captures the variations in the population data. $1(\cdot)$ represents an indicator function used for counting. The overall quality of the model is measured using the F1-score, which is the harmonic mean of these two metrics, as shown in Equation 10.

$$\text{Feasibility} = \frac{1}{M}\sum_{j=1}^{M} 1_{\widehat{X_j}\in X} \tag{8}$$

$$\text{Heterogeneity} = \frac{1}{N}\sum_{i=1}^{N} 1_{X_i\in \hat{X}} \tag{9}$$

$$\text{F1-score} = \frac{2\times \text{Feasibility} \times \text{Heterogeneity}}{\text{Feasibility} + \text{Heterogeneity}} \tag{10}$$

We combined the HTS data collected in 2010 and 2016 from Seoul to create the hypothetical population (h-population). We extracted only the transit trips from the h-population to create the hypothetical SC (h-SC) data and sampled 1% of the data, matching the sampling rate of the HTS, to generate the hypothetical HTS (h-HTS) data.

We compared the proposed data fusion method against benchmark models in Table 1. The partial joint distributions are considered on a trip-chain basis or a trip basis. Prototypical activity schedules refer to the traditional matrix fitting method, which aligns the sample's marginal distribution with the population. VAE-GAN is a single-source model where a single discriminator is connected to a VAE trained solely on h-HTS data. The proposed data fusion method addresses the low heterogeneity problem caused by the low sampling rate of HTS, outperforming the VAE-GAN.

We also conducted a case study that fuses the real-world data sources: the 2016 HTS and SC data collected from Seoul. Figure 1 illustrates the commuting and home return locations observed in the HTS data, as well as those generated by single-source model (No fusion) and CollaGAN models, respectively. The CBD and GBD correspond to Seoul's major business districts. Compared to the VAE-GAN model, the CollaGAN model effectively captured commuting patterns in the business districts while also efficiently generating home-return patterns in the business districts, which are underreported in HTS data.

Table 1 – *Evaluation results of proposed method compared with benchmark models*

Model	Metric	Attribute combination				
		Trip-chain level		Trip level		
		$P(T)$	$P(S)$	$P(T,S)$	$P(T,S,A)$	$P(T,S,M)$
Prototypical activity schedules	Feasibility	100.0%	49.3%	75.7%	71.1%	63.6%
	Heterogeneity	64.4%	7.7%	7.1%	6.7%	3.8%
	F1-score	78.4%	13.4%	13.0%	12.3%	7.3%
VAE-GAN	Feasibility	60.1%	28.1%	67.4%	60.8%	50.0%
	Heterogeneity	80.0%	50.8%	68.9%	62.6%	51.1%
	F1-score	68.6%	36.2%	68.1%	61.7%	50.5%
CollaGAN	Feasibility	77.0%	46.3%	78.1%	72.2%	65.1%
	Heterogeneity	85.3%	68.5%	80.8%	73.7%	60.3%
	F1-score	80.9%	55.2%	79.4%	72.9%	62.6%

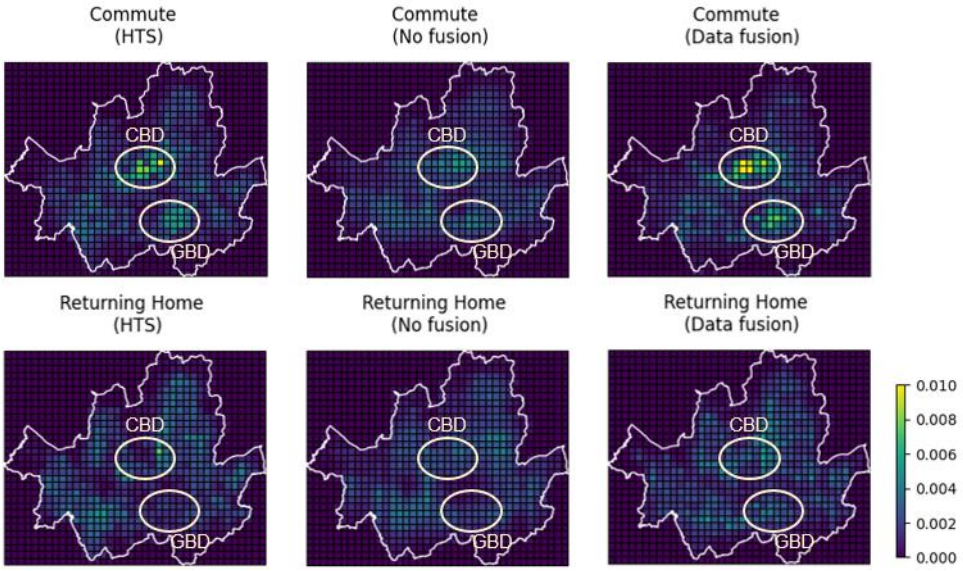

Figure 1 – *Joint distribution of activity purposes and locations in the case study*

Acknowledgements

This work is financially supported by Korea Ministry of Land, Infrastructure and Transport (MOLIT) as Innovative Talent Education Program for Smart City, by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A2C2012835), by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2023-00246523 and No.RS2024-00337956), and by a Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Korean government (MOLIT) (No.RS-2022-001560).

A part of this research was conducted at the Future Cities Lab Global at Singapore-ETH Centre. Future Cities Lab Global is supported and funded by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme and ETH Zurich (ETHZ), with additional contributions from the National University of Singapore (NUS), Nanyang Technological University (NTU), Singapore and the Singapore University of Technology and Design (SUTD). The authors used OpenAI's ChatGPT to correct the typos and the grammar of this manuscript. The authors verified the accuracy, validity, and appropriateness of any content generated by the language model.

References

Borysov, Stanislav S., Rich, Jeppe & Pereira, Francisco C., 2019. How to Generate Micro-Agents? A Deep Generative Modeling Approach to Population Synthesis. *Transportation Research Part C: Emerging Technologies*, **106** pp. 73-97.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron & Bengio, Yoshua, 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, **27**.

Kim, Eui-Jin & Bansal, Prateek, 2023. A Deep Generative Model for Feasible and Diverse Population Synthesis. *Transportation Research Part C: Emerging Technologies*, **148** pp. 104053.

Kingma, Diederik P., 2013. Auto-Encoding Variational Bayes. *arXiv preprint arXiv*: 1312. 6114.