공학박사 학위논문

A Deep Generative Modeling Framework for Synthesizing Individual-specific Activity Schedule

개인별 활동 일정 생성을 위한 심층 생성 모형 프레임 워크

2021년 2월

서울대학교 대학원

공과대학 건설환경공학부

김 의 진

A Deep Generative Modeling Framework for Synthesizing Individual-specific Activity Schedule

지도 교수 김 동 규

이 논문을 공학박사 학위논문으로 제출함
2021년 2월

서울대학교 대학원
공과대학 건설환경공학부
김 의 진

김의진의 공학박사 학위논문을 인준함
2021년 2월

위 원 장 고 승 영 (인)

부위원장 김 동 규 (인)

위 원 이 청 원 (인)

위 원 손 기 민 (인)

위 원 김 진 희 (인)

Abstract

Emerging technologies that provide large-scale high-dimensional travel data have led to the increased use of data-driven approaches for travel demand analysis. Emerging data such as smart card data and locational data from mobile device continuously records mobility trip-chain. However, those data are incomplete due to missing attributes of travel behavior such as trip purpose and individual-specific attributes (i.e., age, gender, and income). Therefore, missing attributes of emerging data need to be estimated for applying them to further analysis, such as travel demand analysis and travel behavior analysis.

Modern advanced activity-based models (ABMs) have adopted the microsimulation approaches whose key inputs are realistic synthetic population (i.e., agent) and activity schedules within a day. Therefore, the data-driven approaches that produce activity schedules of the agent with their individual-specific attributes are necessary to utilize the valuable data collected by emerging technologies. Traditionally, population synthesis and activity scheduling are independently conducted by different approaches. Population synthesis has been performed by estimating the joint probability distribution of individual-specific attributes of populations. Activity scheduling has been conducted by a set of machine learning models for discrete choice, based on numerous expert-designed components. Therefore, a great effort of human experts is required to fit the emerging data by

modifying the existing activity scheduling framework. This study develops a deep generative modeling framework for synthesizing individual-specific activity schedules using mobility data recording trip-chain. The proposed data-driven framework simultaneously estimates the joint probability distribution of activity schedules and individual-specific attributes to generate synthetic activity schedules. The framework can also utilize the incomplete sample data by estimating missing attributes of those data using the estimated joint probability distribution of complete sample data.

To verify my framework, it applied to national household travel survey (NHTS) data (i.e., complete sample data) and transit smart card (SC) data (i.e., incomplete sample data). The estimated joint probability distribution of NHTS data is used to generate realistic synthetic NHTS data (i.e., individual-specific activity schedules) and impute the missing individual-specific attributes of SC data based on a data fusion approach. The synthetic NHTS data and individual-specific SC (ISC) data are evaluated with benchmark models in terms of the probability distribution, the applicability to ABM, and the applicability to behavior analysis. Evaluation results show that the synthetic NHTS well estimate the joint probability distribution of individual-specific activity schedules, and such performances are reasonable for applying to ABM. The results also show that the estimated ISC data represent realistic attributes compared to Census and NHTS data, and it can provide new insights into travel behavior by replacing costly travel survey. These findings indicate that this study can significantly contribute to travel

demand and behavior analysis in terms of data utility, data quality, data privacy, and cost-efficiency.

Keywords: Activity-based Model, Population Synthesis, Activity Scheduling, Deep Generative Modeling, Data Fusion Approach

Student Number: 2016-21248

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Background

Transportation decision-makers should make informed choices for how travel demand changes in response to different inputs considering socio-demographic, economic, and behavioral attributes. Urban transportation systems have begun to provide large-scale high-dimensional mobility data collected from current and emerging technologies such as transit smart-card, GPS-enabled mobile devices, and mobility-as-a-service (MaaS)(Miller, 2017; Sochor et al., 2018). Those data can be used to understand travel behavior for achieving societal goals such as travel demand forecasting and policymaking for reducing traffic congestion. As a remedy, two important issues for utilizing those data have been raised.

One issue is a need for advanced models implemented within a much more flexible framework (Miller, 2017). Existing operational models for travel demand forecasting shared several limitations such as a simplistic representation of travel behavior, poor prediction of destination choice (Miller, 2017; Jung and Sohn, 2017), heuristic rule-based design (Miller and Roorda, 2003; Arentze and Timmermans, 2004), and conditional representation of scheduling (Čertický et al., 2015; Drachal et al., 2019). These limitations are mainly caused by previous attempts that decompose complex modeling tasks into several independent models. While the decomposition of modeling tasks can reduce the model complexity (i.e., the dimensionality of the data) and

consider specific contexts for each task, it requires subject judgments for behavioral assumptions and local calibration of each model for emerging data. Therefore, when the emerging data is introduced, the existing framework based on multiple models requires much effort to modify each model by human-expert with additional validation and calibration. To address this inefficiency, a data-driven framework is needed to conduct complex modeling tasks, minimizing dependence on expert-designed model components.

Another issue is the need to impute missing behavioral attributes of mobility data such as trip purpose and individual-specific attributes (ISA) (i.e., age, gender, and income). For example, while the smart-card collects continuous long-term period of trip-chain of transit, it cannot obtain individual attributes of the smart-card user such as age, gender, and trip purpose. Those missing attributes prevent the emerging data from replacing conventional national household travel survey (NHTS) data, including ISA and activity-trip attributes (ATA) in travel behavior and travel demand analysis. Therefore, missing attributes of emerging data are needed to be estimated for applying emerging data to further analysis of urban mobility.

1.2. Research purpose and scope

Based on the background, this study proposes a methodological framework to address two issues: data-driven framework and missing attributes imputation. For a data-driven framework for emerging data, a data-driven

activity-based model (ABM) is adopted to represent daily travel behavior collected from multi-source data (Drachal et al., 2019). The ABM is constructed under the assumption that travel is a derived demand originating from an individual's need to join the activities. Modern advanced ABMs have adopted the microsimulation approaches whose key inputs are realistic population synthesis and corresponding activity schedules (Arentze and Timmermans, 2004; Roorda et al., 2008; Čertický et al., 2015; Drachal et al., 2019). The general workflows of ABM include three steps: (i) population synthesis, (ii) activity scheduling, and (iii) simulation of executing activity schedule corresponding population. The data-driven ABMs focus on the first and second steps, which defined those tasks as training a generative model (Bishop, 2006). The generative model focuses on discovering the data generation process rather than training the decision-making process by estimating the data's joint probability distribution.

Traditionally, population synthesis and activity scheduling are independently conducted based on different approaches. The population synthesis aim at generating populations described by ISA that affect the choices of activity pattern. Simulation-based approaches that generate representative samples for a given population based on an estimated joint probability distribution have been recently adopted for population synthesis in the previous studies (Farooq et al., 2013; Sun and Erath, 2015; Saadi et al., 2016; Sun et al., 2018; Borysov et al., 2019).

Activity scheduling is a process representing the travel behavior of

the synthesized population. The activity schedule is sequences of activities, including the purpose of activities, location of activities, time and duration of activities, and travel modes to get the activities, which are fundamental input and outputs of ABM. After the 2000s, when a machine learning (ML) model was introduced in earnest, the activity scheduling has been implemented using a set of sequential ML models concerning each choice dimensions such as location, activity duration, activity purpose, and travel mode (Arentze and Timmermans, 2004; Auld and Mohammadian, 2012; Čertickỳ et al., 2015; Drchal et al., 2019). Although the ML models replace many expert-design components based on those flexible structures, existing ABM still suffers challenges from the curse of dimensionality in predicting destination choice (Miller, 2017; Jung and Sohn, 2017) and simultaneously estimating the multiple dimensions of choices. The dimensionality issue has been tackled by the complex decomposition of activity scheduling process based on human cognitive processes (Arentze and Timmermans, 2004; Drchal et al., 2019); however, such approaches still require subject judgments of the expert to design decomposition structure.

As a remedy, I propose a data-driven framework that performs both population synthesis and activity scheduling, based on deep generative modeling (DGM). DGM is a new scalable approach focusing on estimating the joint probability distribution of large-scale high dimensional data. In recent, the DGM that combine a deep neural network and an innovative learning algorithm (Kingma and Welling, 2013; Goodfellow et al., 2016) have

succeeded in training data generation process of high-dimensional data such as music (Yang et al., 2017), images (Van den Oord et al., 2016), and natural language (Wen et al., 2015). I extend the progress of DGM into ABM that models complex interrelation between discrete choices for activity scheduling and representative individual attributes. The proposed framework re-generates high-dimensional activity schedules and corresponding individual attributes by estimating the joint probability distribution of the mobility trip-chain data using DGM. Those generated data (i.e., synthetic data) can be applied to the ABM as key inputs and be provided to the researcher to investigate travel behavior without concerns of data privacy since the data do not directly correspond to real people's activities.

For missing attributes imputation, the DGM is also used based on a data fusion approach. The data fusion approach aims at integrating multiple data sources based on the assumption that relationships between attributes of complete sampled data are similar to those of incomplete sampled data. The missing attributes of large-scale incomplete data can be estimated by a probability distribution of missing attributes observed in the small-scale complete sampled data. In transportation research, several studies have applied the data fusion approach to estimate trip purpose and/or travel mode of the smart card and/or GPS track data (Kusakabe and Asakura, 2014; Xiao et al., 2015; Bantis and Haworth, 2017; Yazdizadeh et al., 2019; Kim et al., 2020). Their studies estimated a conditional probability of the single missing attribute (i.e., trip purpose or travel mode) in the complete sampled data using

the ML model. The estimated probability distribution is used to deduce the single missing attributes in the incomplete sampled data.

This study proposes an extension of the data fusion approach using conditional DGM to estimate the incomplete data's missing attributes. While an unconditional DGM has no control attributes of the generated data, the conditional DGM can generate the data for given observed attributes. Such conditioning could be based on class labels, some part of data, and even data from a different source (Mirza and Osindero, 2014). The proposed conditional DGM is used to estimate the ISA of smart-card (SC) data (i.e., incomplete sample data). The conditional DGM are trained to generate the ISAs given ATAs using NHTS data. While the ATAs are observed in both NHTS (i.e., complete sample data) and SC data, the ISAs are only included in NHTS data. By applying the conditional DGM to SC data, missing individual attributes of SC data are estimated. These generated individual-specific SC (ISC) data can represent interesting travel behaviors in urban mobility, such as an individual's long-term or specific travel behaviors, which cannot be obtained from the existing single data.

The proposed framework consists of three modules: Generation module (GM), merging module (MM), and labeling module (LM). The GM and MM address the first issues as a data-driven ABM. The GM is trained to estimate the joint probability distribution of high-dimensional complete sample data using DGM. NHTS data collected in the Seoul metropolitan area are used as complete sampled data in this study. The GM is trained to generate

the individual-specific consecutive trips in Seoul, i.e., synthetic NHTS data of consecutive trips, rather than the entire trip-chain. The NHTS data only collect the individual's travel diary in a single day, making the extraction of trip-chaining patterns in a data-driven manner difficult. The MM complements the GM by creating an individual-specific trip-chain using the generated consecutive trips. An efficient trip-chaining algorithm based on the estimated joint probability of GM is proposed to merge the synthetic consecutive trips. The algorithm of MM is designed based on the observed travel patterns in the data. The outputs of MM and GM are synthetic NHTS data of trip-chain. The performance of GM and MM are evaluated with the state-of-the-art benchmark models in terms of the probability distribution and the applicability to ABM.

The LM addresses the second issue by estimating missing attributes of incomplete sample data. The joint probability distribution of ISAs given ATAs is estimated from NHTS data using conditional DGM, and it is used to impute missing attributes of SC data that are used as incomplete sampled data in this study. This task is referred to as a "labeling" of which missing attributes of the incomplete data are estimated. The outputs of LM are ISC data of trip-chain. The performance of LM is evaluated using multi-source data in terms of probability distribution and the applicability to travel behavior analysis. Also, applications of LM to the analysis of monthly travel behavior and customized travel behavior are proposed to verify the rationality and utility of the generated ISC data.

Figure 1 represents the overview of the proposed framework associated with three modules and the corresponding dataset. The complete sample data, NHTS, and the incomplete sample data, SC, represent and evaluate the proposed framework. The GM and MM re-generate the synthetic NHTS data, and the LM generates the ISC data. These generated data have four strengths compared to the original data, such as NHTS and SC data: data quality, data privacy, data utility, and applicability to ABM. In terms of the data quality, the synthetic data can generate the missing activities from the original sampled data but exist in the population data. These out-of-sample activities can be used to represent a realistic set of modeling results. In terms of data privacy, the synthetic data do not directly correspond to real people's mobility records such that valuable individual-specific micro-mobility data can be provided as a public dataset. In terms of data utility, a finer and specific travel demand analysis can be conducted using the ISC data. Lastly, since the ABM requires the agent-based sample generated by the probability distribution of data, the output of the proposed framework can be directly used as input for ABM. Synthetic NHTS data have strength over NHTS for data quality, data privacy, and applicability to ABM, and ISC data have advantages over SC for data utility.

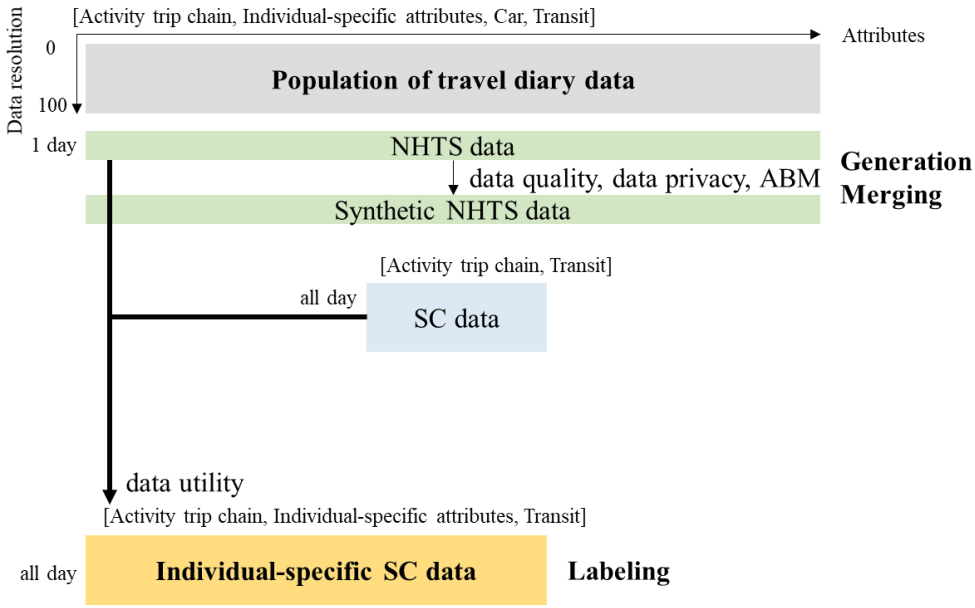


Figure 1. Overview of the proposed framework associated with three modules: Generation, Merging, and Labeling.

1.3. Research contribution

The main contributions of this paper are described as follows. First, an integrated framework for generating individual-specific activity schedules is developed by connecting the area of ABM with DGM. This study represents how two widely used DGMs can be tailored to generating individual-specific activity schedules. Second, a data-driven framework is developed to integrate population synthesis and activity scheduling using a single DGM (i.e., GM) and an efficient algorithm (i.e., MM). The framework re-generates the synthetic NHTS data (i.e., complete sample data) by estimating the joint probability distribution of individuals' consecutive trips and merging these consecutive trips. Third, a data fusion approach is extended to estimate

multiple missing attributes using conditional DGM, which combines the NHTS and SC data to complement the quality and utility of the data. Fourth, the modules making up the framework are evaluated by comparing with state-of-the-art benchmark models in terms of the probability distribution, the applicability to ABM, and the applicability to behavior analysis. Gibbs sampler using the classification and regression tree (GCART) (Drechsler and Reiter, 2011), Wasserstein generative adversarial network with gradient penalty (WGAN-GP) (Gulrajani et al., 2017), Variational autoencoder (VAE) (Kingma and Welling, 2013) were compared or combined to find the best model for generating large-scale travel data. Fifth, the evaluation results show a strong promise to generate large-scale travel data using complete and incomplete trip-chain data. These results show that the synthetic NHTS data represent the data generating process of individual-specific activity schedules, and such performance is reasonable for applying to ABM. The LM estimated realistic ISA of SC data, and those estimated ISA are evaluated using Census and NHTS data. The ISC data represent not just a realistic probability distribution of ISA but also provide new insights into travel behavior, which can replace costly travel survey. These findings indicate that this study can significantly contribute to travel demand and behavior analysis in terms of data utility, data quality, data privacy, and cost-efficiency.

The remaining part of this study is organized as follows. In Chapter 2, related works concerning ABM, DGM, and the data fusion approach are extensively reviewed, and their advantages and shortcomings are examined.

In Chapter 3, the hypothesis, flowchart, and data used in the study are described in detail. Chapter 4 describes the details of the method in each module, and Chapter 5 shows evaluation results with the benchmark models in terms of the probability distribution, the applicability to ABM, and the applicability to travel behavior analysis. Lastly, in Chapter 7, I discuss my findings, make concluding remarks, and share future research plans with the limitations of this study.

Chapter 2. Related Works

2.1. Activity-based Model

Activity-Based Models (ABM) (Ben-Akivai et al., 1996) are a specific agent-based model for travel demand analysis. The ABM simulates the individual activities of agents and their impact on the transportation system's operation to understand and predict future travel demand. Simulating individual activities of agents are implemented using the probability distribution of individual activities estimated using complete travel data such as NHTS data. Compared to the traditional trip-based model (i.e., 4-step model), ABM can represent the interrelations among activities, travel patterns, and person (Castiglione et al., 2015). These activity-related travel behaviors can be described as activity schedules (i.e., activity trip-chain) that consisted of activity start time and duration, activity location, activity type (e.g., work, shopping, and home), and travel mode to get the activity. Generally, the individual activity schedules are represented as daily interconnected trips starting and finishing at home within a day.

Most activity-based models generally incorporate three essential steps (Catiglione et al., 2015): (i) population synthesis that generates agents' comprehensive individual attributes; (ii) activity scheduling that simulates choices of individuals about daily activities; (iii) network assignment that assigns the generated activity schedules as travel demand on the transportation networks and estimates the level-of-service of the networks.

This study focuses on the first and second steps, where the synthetic population with activity schedules are generated.

Two methodological perspectives have been mainly proposed for the population synthesis: (i) re-weighting and (ii) simulation-based approaches (Tanton, 2014). The re-weighting approach aims at estimating weighting factors for a given sample in the travel data, and the re-weighted samples are used as representatives of the population. Re-weighting is conducted by matching specific attributes such as home location to macroscopic constraint data such as census (Hermes and Poulsen, 2012). This task is typically achieved by applying non-linear optimization to estimate weights. Matrix fitting is a specific kind of re-weighting approach that estimates expansion factors. Popular methods for matrix fitting are Iterative Proportion Fitting (IPF) (Deming and Stephan, 1940) and maximum cross-entropy (Guo and Bhat, 2007). The matrix fitting method estimates a multivariate table of individual attributes as expansion factors, and the population is sampled from the complete sample to match the expansion factors. Since the re-weighting approaches generate a weighted individual sample from the complete sample data, rather than an agent-based sample from the probability distribution. Therefore, to apply re-weighting approaches to ABM, the simulation stage is conducted by re-sampling from the weighted samples. The simulation-based approaches have been received increasing attention since it can provide a more systematic way of generating data. Even if specific agents do not exist in the original data, it may still be possible to sample these specific agents

using the estimated probability distribution. Also, simulation-based methods often have the advantage of being effective for high-dimensional problems, which address more detailed populations.

The simulation-based approach aims at estimating joint probability distribution for a given sample in the travel data. A Bayesian network that models a joint probability distribution of the full set of variables (i.e., attributes of agents) is proposed for simulation-based population synthesis by Sun and Erath (2015). The authors argue that the Bayesian network can effectively make the inference if the conditional structure of the data generating process is known. However, as this conditional structure is typically not available, they propose that the data's graph structure is learned through a scoring approach. Although their model performs well against IPF and a Gibbs sampler for small dimensions of sample, the graph structure's learning task is known to be a computationally challenging task, which suffers the scalability issue. Sun et al. (2018) proposed a customized hierarchical mixture modeling framework for population synthesis. Latent variables defined by individual-specific attributes constitute the hierarchies of the model. Based on these hierarchies, the joint probability distribution of all the attributes is estimated as the product of all the marginal distributions of the variables conditioned on the hierarchical latent variables. Expectation-Maximization (EM) algorithm is used to estimate the model parameter, and the model can capture both the marginal and combined distributions of all the variables using the latent variable representation. However, this approach still

suffers the scalability issue to many dimensions and robustness to the definition of the latent variables. Saadi et al. (2016) proposed a Hidden Markov Models (HMMs) to model the joint probability distribution of the population's attributes. The key idea is that the observed attributes representing certain phenomena are associated with hidden states of those phenomena. The hidden states are fully determined by their previous state (i.e., a Markovian process), and this relationship is measured by the transition matrix (i.e., probability matrix determining the transition rates between states). All the attributes are generated in sequences from the HMM to represent all attributes for a given individual. However, defining the natural ordering of attributes to define the chain of states is also computationally expensive, which causes the scalability issue.

The representative studies of early ABM models are TRANSIMS (Smith et al., 1995), CEMDAP (Bhat et al., 2004), and SacSim (Bradley et al., 2010). These models are mainly constructed using heuristic rules designed by human-expert. In recent, the modern ABM have tried to complement those heuristic rules by defining them as ML problem. The ALBATROSS model (Arentze and Timmermans, 2004) and the Toronto Area Scheduling model for Household Agents (TASHA) (Miller and Roorda, 2003) were two notable works that generated hourly activity schedules and travel patterns for a typical weekday. ADAPTS (Auld and Mohammadian, 2012) is another activity scheduling algorithm focusing on the ordering and timing of individual scheduling decisions. Čertickỳ et al. (2015) developed a sophisticated ABM

applied in European Cities. Their approach was also implemented using both the ML methods and expert knowledge. All these aforementioned models rely on heuristics generated by human-experts, which require a great effort to model maintenance and deployment.

More recently, data-driven approaches have been proposed to leverage the increasing availability of emerging data that represent valuable travel information. The data-driven ML methods that estimate the probability distribution of the data were used to model multiple individual choices in activity scheduling, which represent the travel behavior of the modeled population. This data-driven ML method has several strengths, such as the capability of modeling complex decision processes, the flexibility of the model for a new location, and easy-to-use implementation with a small team for maintenance and management. Simple ML methods such as logistic regression model, linear regression model, and decision tree are applied to activity scheduling (Arentze and Timmermans, 2004; Bhat et al., 2004; Auld and Mohammadian, 2012). More advanced ML methods such as tree-based ensemble methods (Rasouli and Timmermans, 2014; Hafezi et al., 2018) and generative neural network models (Čertickỳ et al., 2015) have begun to proposed in recent studies. The ML methods are often used to model the subcomponents of activity scheduling, such as activity type choice, activity duration choice, activity location choice, and travel mode choice. Drchal et al. (2019) proposed a data-driven ABM framework by designing an activity scheduling problem as a generative modeling task. Their model significantly

reduces the expert-designed components by extracting key rules of activity scheduling components using the estimated probability distribution of the data.

2.2. Deep Generative Model

Recently, DGMs have been reported their excellent performance and computational effectiveness in generating high-dimensional data such as images, natural language, and tabular data (Yang et al., 2017; Van den Oord et al., 2016; Wen et al., 2015; Xu et al., 2019). Well-known DGMs are VAE (Kingma and Welling, 2013) and GAN (Goodfellow et al., 2016). These DGMs re-generate the statistical representations of the original data by sampling the estimated joint probability distribution of the data. While GANs have been used extensively for the image, sound, and sequential text generation, VAEs have been applied to not just image and text but also the tabular data that is composed of both numerical and discrete attributes (Borysov et al., 2019).

Transportation researchers have tried to use these outstanding performances of DGM in the transportation modeling task (Barthlemy and Carletti, 2017; Drchal et al., 2019). DGMs were one of the most promising approaches to the high-dimensional modeling task in transportation planning, such as extracting users' travel patterns (Yin et al., 2018) and generating synthetic travel data (Lin et al., 2017) by estimating the statistical properties

of real people's travel behavior. In particular, Borysov et al. (2019) reported that a vanilla VAE shows excellent performance in synthesizing population from large-scale travel data by estimating the joint probability distribution of the original data. Badu-Marf et al. (2020) proposed the composite GANs to estimate the joint probability distribution of travel data having individual-specific attributes and sequential destination locations. The authors showed that their customized structure of GANs outperformed the VAE in generating travel diary data. Lin et al. (2017) developed a modeling framework that processes the locational cell phone data to synthesize activity trip-chain data. Input-output hidden Markov model (IO-HMM) was used to estimate the trip purpose of the locational data in an unsupervised manner, and DGM based on long short term memory (LSTM) were used to learn joint probability distribution of activity trip-chain. Synthetic activity trip-chain data were then evaluated using observed traffic and transit passenger counts.

2.3. Data fusion approach

Data fusion is one of the approaches to integrate multiple data from different sources. Given target attributes that are not collected from incomplete data but complete data, statistical learning methods estimate the conditional probability distribution of the target attributes given observed attributes of incomplete data. For example, Shen and Stopher (2013) developed a trip purpose estimation method for locational data by using NHTS data. In their

method, the trip purposes which were not directly observed by locational data were estimated using rules obtained from the NHTS data. Kusakabe and Asakura (2014) also employ the data fusion methodology to derive relationships among behavioral attributes that cannot be obtained from either smart card data or survey-based data alone. They apply the naïve Bayes classifier to estimate the probability distribution of trip purpose in NHTS data, given activity trip attributes in smart card (SC) data. Bantis and Haworth (2017) proposed dynamic Bayesian networks that use the environmental and individual attributes as a prior distribution to estimate the travel mode of locational data. Their results revealed a significant impact of individual attributes on travel mode choice, which would inform policymakers into actions targeted at marginalized population groups. Kim et al. (2020) proposed an interpretable machine learning approach to apply the data-fusion approach to behavioral analysis. The authors showed that interpretation methods for ML could reveal the dominant factors in estimating the trip purpose of SC data, which provides insight to construct the complete sampled data.

Chapter 3. Problem Definition

3.1. Probabilistic design

A detailed description of GM and LM are described by formulating generative modeling in this subsection. The GM learns the joint probability distribution of NHTS data that consists of ISA ($\boldsymbol{u}$), consecutive activity trips ($\boldsymbol{a_i}$,$\boldsymbol{a_{i-1}}$). $\boldsymbol{a_i}$ is the activity trip attributes (ATA) of *i*-th trip representing trip purpose, activity duration, destination, and travel mode, and $\boldsymbol{a}$ is the ATA of an activity trip-chain that is a set of all activities within a day. The ATAs observed in the NHTS data ($\boldsymbol{a^{NHTS}}$) are differentiated from those in the SC data ($\boldsymbol{a^{SC}}$) since those data populations could be different. While the NHTS data are survey data obtained from 1-2% of the population, the SC data are revealed preference data obtained from 100% of the transit users. The $\boldsymbol{a^{NHTS}}$ and $\boldsymbol{u}$ generated by estimated probability distribution are marked as $\widehat{\boldsymbol{a^{NHTS}}}$ and $\hat{\boldsymbol{u}}$, respectively. Table 1 describes the input, modeling task, and output of the modules in the proposed framework. The input and output of the LM are different in the training and test process (i.e., the data-fusion approach) under the assumption that interrelations between ATA and ISA are the same in NHTS and SC data. To conduct the modeling tasks in Table 1, various benchmark models and algorithms are evaluated to determine the best model. This effort is described in Chapter 5.

Table 1 Input, modeling task, and output of the modules making up the proposed framework.

Module		Input	Modeling task	Output
Generation		$\boldsymbol{a_i^{NHTS}, a_{i-1}^{NHTS}, u}$	$P(\boldsymbol{a_i^{NHTS}, a_{i-1}^{NHTS}, u})$	$\widehat{\boldsymbol{a_i^{NHTS}}}, \widehat{\boldsymbol{a_{i-1}^{NHTS}}}, \boldsymbol{\hat{u}}$
Merging		$\widehat{\boldsymbol{a_i^{NHTS}}}, \widehat{\boldsymbol{a_{i-1}^{NHTS}}}, \boldsymbol{\hat{u}}$	Trip-chaining	$\widehat{\boldsymbol{a^{NHTS}}}, \boldsymbol{\hat{u}}$
Label	Training	$\boldsymbol{a_i^{NHTS}, a_{i-1}^{NHTS}, u}$	$P(\boldsymbol{u}\|\boldsymbol{a_i^{NHTS}, a_{i-1}^{NHTS}})$	$\widehat{\boldsymbol{a_i^{NHTS}}}, \widehat{\boldsymbol{a_{i-1}^{NHTS}}}, \boldsymbol{\hat{u}}$
	Test	$\boldsymbol{a^{SC}}$		$\boldsymbol{a^{SC}, \hat{u}}$

The output of GM and MM can be used as an input of ABM that is a modern travel demand forecasting approach. The probabilistic design of the proposed framework is compared with the existing ABM and 4-step model to clarify the differentiation, as shown in Table 2. Based on the main tasks of the ABM (i.e., population synthesis and activity scheduling), they share the matrix fitting steps where the sampled are expanded to represent the estimated populations. The proposed framework cannot cover the matrix fitting stage, but more details of matrix fitting in the transportation research are described in Pritchard and Miller (2012) and Choupani and Mamdoohi (2016). In the starting solution of population synthesis, this study and existing ABM generates population from an estimated joint probability distribution of NHTS while the 4-step model used the raw NHTS. The use of raw NHTS data causes the problem of sampling zeros, that is, the agents that are missing from the NHTS data but exist in the real population. The 4-step model does not consider the individual activity patterns and just forecast the aggregated travel demand in daily and TAZ units in the activity scheduling. The existing ABM can predict finer travel demand based on the detailed person-level

analysis. It represents activity scheduling as sequential discrete choices of long-term mobility (e.g., car owner, home, and work location), activity destination, activity duration and time, and access travel mode. Those different choice dimensions are decomposed by ML or parametric models to reduce the model complexity and deal with different data in each choice dimension. In contrast, the proposed framework integrates the activity scheduling and population synthesis by estimating the full-joint probability distribution of complete activity trip-chain data based on the outstanding performance and computational effectiveness of DGM. In particular, the proposed framework based on DGM would outperform for monitoring travel behavior according to the changes of travel pattern since it fully estimates probability distribution of current travel behavior without any parametric assumption or decomposition of the decision making process.

Table 2 Comparison of the proposed framework with existing ABM and 4-step model in the process of population synthesis and activity scheduling

<table>
<tr><th>Tasks</th><th>Stages</th><th>This study</th><th>Existing ABM</th><th>4-step model</th></tr>
<tr><td rowspan="2">Population synthesis</td><td>Starting solution</td><td>Joint probability distribution of NHTS</td><td>Joint probability distribution of NHTS</td><td>NHTS (fixed one day)</td></tr>
<tr><td>Expansion</td><td colspan="3">Sample weights are calibrated using matrix fitting</td></tr>
<tr><td rowspan="4">Activity scheduling</td><td>Long-term</td><td rowspan="4">Joint probability distribution of NHTS</td><td>Long-term choice</td><td>Trip generation</td></tr>
<tr><td>Destination</td><td>Destination choice</td><td>Trip distribution</td></tr>
<tr><td>Time-of-day</td><td>Time-of-day choice</td><td>-</td></tr>
<tr><td>Travel mode</td><td>Mode choice</td><td>Mode choice</td></tr>
</table>

3.2. Hypothesis

Three modules in the proposed framework are constructed based on several hypotheses for sampled and population of travel data. Table 3 summarizes the hypothesis and corresponding limitation applied to this study in terms of data and modeling. First, the proposed framework assumes that NHTS data represent the travel pattern of the population. Specifically, while the traditional 4-step model assumes the NHTS data at the time of survey represent the population, the proposed framework assumes that the probability distribution of NHTS represents the population of travel diary data better than NHTS since missing data in the NHTS can be generated by the estimated probability distribution. Low sampling rate (i.e., 1.13 % in Seoul) and response bias of survey participants such as a cognitive error on activity duration and exclusion of minor trips can be limitations of this hypothesis. Second, this study assumes that the interrelations among ATAs and ISAs do not change; thus, the estimated probability distribution of NHTS data at a specific time can be applied to future travel demand forecasting. This hypothesis is also applied in the existing ABM (Castiglione et al., 2015), which is a relaxed assumption compared with the 4-step model with an invariant assumption of trip distributions (i.e., OD trips). Third, this study assumes that SC data represent all travel patterns of transit users since all transit uses in Seoul are recorded in SC data. However, people who use the transit can also use the other travel modes such as taxi, bike, and private car

and travel of those people would be missing. I investigated the NHTS data to figure out the proportion of people using both transit and other travel modes, and 2.49% of people used those multi travel modes within a day. Forth, I hypothesized that the joint distribution of NHTS represents the population better than raw NHTS. The population of NHTS data does not exist at the moment; thus, the SC data that are the population data of transit users is used to verify the hypothesis for transit users only. The verification results are provided in Chapter 6. Lastly, this study assumes that the interrelation between ATA and ISA does not change in the NHTS and SC data, which is the core assumption for LM's data-fusion approach. The LM estimates the ISA of SC data, but the population of those data (i.e., ISA of transit users) does not exist. Using census data that are the population of ISA of all travelers, the estimated ISA is indirectly validated.

Table 3 Summary of the hypothesis in this study

Perspective		Hypothesis	Limitation
Data	NHTS	・NHTS data represent the travel pattern of the population	・Low sampling rate (1.13%) ・Response bias
	NHTS	・Interrelation among ATAs and ISAs does not change.	・Interrelation can change due to a new travel mode or a new town
	SC	・SC data represent all travel pattern of transit user.	・People who use both transit and other travel modes (2.49%)
Method	GM	・The joint distribution of NHTS better represent population than NHTS.	・Only transit user data can be validated using SC data.
	LM	・The interrelation between attributes is the same in NHTS and SC.	・Only indirect validation can be conducted using Census data.

3.3. Data description

3.3.1. NHTS data

The NHTS data records the individual travel diary in a single day, including the ATA, such as trip purpose, activity destination, activity duration, travel mode, and ISA such as age, gender, income, household size, type of home. Although the NHTS can provide complete information for travel demand and travel behavior analysis, monitoring recent travel patterns is limited because the NHTS is conducted every five years due to the large costs required for the survey.

This study used the NHTS dataset collected in 2016 (KTDB, 2016), and only the data in Seoul were used to match the spatial area with the SC data. These data included individual travel diaries that recorded every daily trip taken, with multiple trips on a given day expressed as a trip chain. I divided the chained trips by their trip purpose and established the major travel modes of the trip's purpose. For example, a person who uses the subway to go to work must first access the subway station on foot and then use the subway. In this case, the two chained trips, walking and subway, are combined into one subway trip as the primary travel mode. Walking is considered a primary travel mode only if used as the sole travel mode, but not as a means to access another travel mode. Seoul operates a public transit unified fare system for buses and subways, whereby charges are levied as if they are using a single travel mode when transferring between these two forms of public

transit. Therefore, this study makes no distinction between a bus and a subway, whereby the chained trips of a bus and subway with a transfer are considered one trip by public transit.

When the joint probability distribution of NHTS data is estimated in the GM, the trip-chain of NHTS data is transformed into consecutive trips to reflect the travel context of the trip chain. Although estimating the distribution of the full trip-chain could be possible if the activity-trip chain data representing long-term chaining pattern exists, this study only can utilize the NHTS data that records a trip-chain only for a single day.

The departure and arrival locations of NHTS data are recorded in the TAZ unit, which is within a radius of about 1 km. Estimating activity destinations is a well-known challenging problem due to its high-dimensionality (Rasouli and Timmerman, 2014). For example, there are 424 TAZ in Seoul, which indicates that modeling departure and arrival locations require modeling 424+424 dimensions of binary data. To reduce the dimensionality of the departure and arrival locations, the locational data collected in the TAZ unit are converted into 19 latitude and 19 longitude data (i.e., 19×19 grids). In other words, modeling tasks of 424 dimensions of two variables are reduced to 19 dimensions of four variables. Both TAZ and coordinates formats were tested in modeling GM and LM. While the GM that considers high-dimensional data performed better for the coordinate form of locational data, the LM that models relatively low-dimensional data showed better performance for the TAZ format of locational data.

The duration of an activity is calculated by the difference between the arrival time on the previous trip and the departure time on the next trip. The duration of activity on the last trip (i.e., the return trip home) is calculated by the difference between the arrival time of the last trip and the first trip's departure time. The travel time includes in-vehicle and out-of-vehicle time, such as waiting time and access time. The number of trips includes all the trips made during a day. Travel modes are divided into a car, walking, transit, taxi, and bicycle. Trip purposes are categorized into a home, work, and others (e.g., leisure and shopping). Individual attributes include age, gender, car availability, driver's license, and household income, and all of those attributes are directly collected in the NHTS data.

Estimating the exact ATA or ISA is meaningless in the dimension of transportation planning or policymaking; rather, estimating the probability of which a traveler will have certain ATA and ISA under certain conditions are enough. All the numerical attributes are converted into categorical attributes of 19 uniform bins to be used as an efficient dimension for the modeling task. Note that discretization of numerical attributes can vary depending on the desired resolution of the travel demand or behavior analysis.

After a data-cleaning process, we removed the trips with very long activity duration and travel time, a total of 180,438 trips taken by 78,761 individuals were used for modeling GM. In modeling LM, only the NHTS data of transit users are used to match the distribution of the NHTS data with those of the SC data. A total of 64,634 transit trips taken by 32,782 individuals

were used for modeling LM. In both GM and LM, I used 50% of the data for training, 20% of those data for validation, and 30% of those data for the test.

Table 4 describes the ISA and ATA collected in NHTS. All the attributes in the table are included in the synthetic NHTS data generated by GM. The fourth column indicates whether the attributes are also collected in SC data or not, and the fifth column classifies the estimated attributes using the LM, and the input attributes of LM collected from SC data. The categories of all variables are taken from NHTS data, and some variables in the SC data are transformed into the form of NHTS data.

Table 4 Description individual-specific and activity trip attributes collected in NHTS data

Attributes	Type(dimension)	Remarks	SC	LM
Household income	Categorical(6)	< 1 million won; 1~2 million won; 2~3 million won; 3~5 million won; 5~10 million won; > 10 million won	X	Output
Car availability	Categorical(2)	Yes or no	X	Output
Driver's license	Categorical(2)	Yes or no	X	Output
Gender	Categorical(2)	Man; Woman	X	Output
Home type	Categorical(6)	Apartment; Villa; Multi-family; Single house; Officetels; Others	X	Output
Student	Categorical(2)	Yes or no	O	Input
Number of household	Categorical(7)	1 ~ 7	X	Output
Age	Numeric(19)	About 5 years interval	X	Output
Home latitude	Numeric(19)	About 1.4 km interval	X	-
Home longitude	Numeric(19)		X	-
Work latitude	Numeric(19)		X	-
Work longitude	Numeric(19)		X	-
Trip purpose (t)	Categorical(3)	Commute; Other; Returning home	X	Output
Trip purpose (t-1)	Categorical(3)		X	Output
Travel mode (t)	Categorical(5)	Private car; Bike; Transit; Taxi; Walk	X	-
Travel mode (t-1)	Categorical(5)		X	-
Destination latitude (t)	Numeric(19)	About 1.4 km interval Format for the GM	O	-
Destination latitude (t-1)	Numeric(19)		O	-
Destination longitude (t)	Numeric(19)		O	-
Destination longitude (t-1)	Numeric(19)		O	-
Destination TAZ (t)	Categorical(424)	424 TAZ in Seoul Format for the LM	O	Input
Destination TAZ (t-1)	Categorical(424)		O	Input
Activity duration (t)	Numeric(19)	About 50 min interval	O	Input
Activity duration (t-1)	Numeric(19)		O	Input
Trip sequence (t)	Categorical(6)	1 ~ 6	O	Input
Last trip (t)	Categorical(2)	Yes or no	O	Input
Travel time (t)	Numeric(19)	About 15min interval	O	Input
Number of trips	Categorical(5)	2 ~ 6	O	Input
Start-time	Numeric(19)	7 ~ 24	O	Input

3.3.2. Smart card data

The smart card system automatically and continuously collects every transit passenger's origin, transfer, and destination stop with exact time-stamps in the Seoul area. However, compared with the NHTS data, the smart card data did not collect individual-specific contexts such as trip purpose and individual-specific attributes. Also, since trips by other travel modes consisting of daily trip-chain are not recorded in the SC data, it could only provide a partial travel pattern of people who use other travel mode within a day. The fourth column of Table 4 indicates the behavioral attributes collected from the SC data, and those attributes also can be collected from NHTS data; thus, the data fusion approach can be applied to the NHTS and SC data. As with NHTS data, all the numerical variable was converted into categorical attributes of 19 bins. The trip purpose of each trip is estimated by the model that learns conditional probability of trip purpose given attributes of SC data, developed by Kim et al. (2020). More details about the trip purpose estimation model are presented in Kim et al. (2020).

Activity durations calculated by the difference between the arrival time at the prior station and the departure time at the next station were reported as important attributes in determining activities (Alsger et al., 2018; Kusakabe and Asakura, 2014). Therefore, the trip-chain with more than two trips within a day is included in the SC data. After a data-cleaning process, in which we removed the trip-chain with more than six trips, very long activity duration, and travel time, a total of 6,247,992 trips taken by 2,690,274

individuals on Nov 16, 2016, were used to evaluate the LM. The application of the generated ISC by the LM was investigated in Chapter 6. In this application, 28 days of SC data from Nov 1 ~ Nov 28, 2016, were used to investigate regular transit users' monthly travel behavior (i.e., people who traveled more than 3 days during a month). Due to the computational complexity, only 10% of regular transit users (i.e., 461,859 peoples in Seoul) were investigated. Interesting descriptive statistics of long-term travel behavior of transit users are provided in Chapter 5.

Chapter 4. Methods

4.1. Flowchart of the framework

The proposed framework aims to synthesize individual-specific activity schedules based on three interrelated modules: GM, MM, and LM. The flowchart of the framework is depicted in Figure 2. The GM uses the NHTS data as an input (i.e., complete sample data) to estimate the full joint probability distribution of NHTS data. The NHTS data are transformed into the individual-specific consecutive activity trips, and the joint probability distribution of the consecutive activity trips is estimated using DGM. The synthetic NHTS data of consecutive trips are then processed to make up the synthetic NHTS data of activity schedules (i.e., activity trip-chain) by the MM. These complete synthetic NHTS data can be applied to travel demand analysis, travel behavior analysis and re-used to model validation and calibration. Meanwhile, the LM uses the SC data as an input (i.e., incomplete sample data). The missing attributes, such as trip purpose and individual-specific attributes, are labeled by the LM that learn the probability distribution of ISA given ATA from NHTS data using conditional DGM.

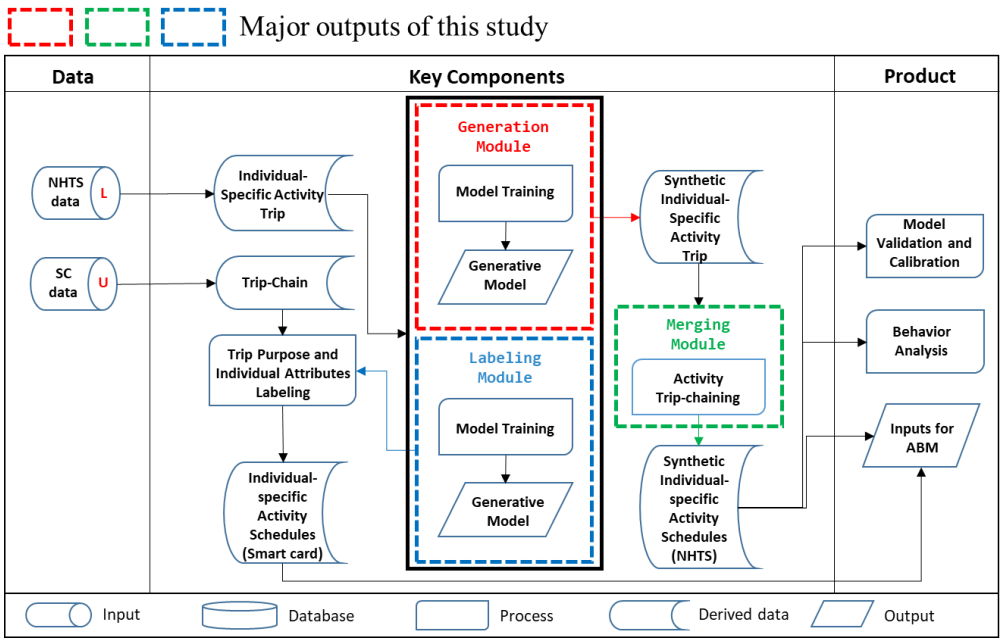


Figure 2. Flowchart of the proposed framework for synthesizing individual-specific activity schedules: generation module, merging module, and labeling module

4.2. Deep generative models

Deep learning has recently witnessed an outstanding success in transportation research by enhancing the accuracy and robustness of the existing models. Wang et al. (2019) provided a holistic literature review on how deep learning can make the transportation system more intelligent. They categorized the applications in the transportation research that rely on an accurate learning model into visual recognition tasks, traffic speed prediction, and travel risk prediction, and traffic signal control. With advanced model structures such as a convolutional neural network (CNN) and recurrent neural network (RNN), those achievements are mainly obtained from a deep discriminative model

that aims to decision making for given conditions of the data. In recently, DGMs have been introduced for solving specific problems in transportation research such as traffic state estimation (Liang et al., 2018), traffic data generation (Lv et al., 2018), and travel time estimation (Zhang et al., 2019). The DGM is another aspect of deep learning that aims to re-generate (i.e., synthesize) the data by estimating the joint probability distribution of the data. Unlike the decision-making achieved by the discriminative model that estimates conditional probability distribution, the DGM pursued an understanding of the data to achieve more general applications such as finding underlying factors, predicting future conditions, finding analogies, and also decision making.

Scalability problems that limited the use of a model to low-dimensional problems or force to simplifying assumptions need to be addressed in generative modeling for data-driven ABM since the individual-specific activity schedules are large-scale high-dimensional data. The DGMs that combines the strength of generative models and deep learning achieved great success in generating high-dimensional data. Two main approaches of DGM are GAN and VAE, which are latent variable models. The latent variable models assume that the data ($\mathbf{x}$) is lying on a low-dimensional manifold (p_λ) and train the generator (G_θ) as a $\mathbf{x} = f_\theta(\boldsymbol{z})$ where the latent variable $\mathbf{z}$ is following probabilistic distribution (i.e., prior over latent space) of $\mathbf{z}{\sim}p_\lambda(\mathbf{z})$. GAN and VAE train this generator with different approaches. While the GAN aim to achieve an equilibrium between the generator that tries

to fool the discriminator by generating realistic data and the discriminator that tries to distinguish between the real and generated data, the VAE aim to maximize the lower bound of the data log-likelihood.

The GANs and VAE have their own strengths and weaknesses in three different perspectives: data quality, training stability, and applicability. For the data quality, VAE tends to spread probability mass to places, which cause the edge blurring in the image data, while the GANs suffer missing mode problem that causes ignoring minor class of the data. For the training stability, while the VAE can be stably trained by optimization-based convergence with tractable likelihood, the GANs suffer unstable training due to intractable discriminator loss (i.e., the quality of data is not proportional to the discriminator loss) and equilibrium-based convergence. For the applicability, the GANs has difficulty in generating discrete data since the adversarial loss based on Kullback–Leibler (KL) divergence requires the continuous output data of the generator for backpropagation. In contrast, the VAE should assume the continuous latent variable for training networks using the backpropagation algorithm, but it works well for discrete data (Borysov et al., 2019). Some modifications of VAE and GANs for discrete data generation have been proposed in the previous studies, such as categorical reparameterization with Gumbel-softmax for a discrete latent variable of VAE (Jang et al., 2016) and Wasserstein distance (Gulrajani et al., 2017) that replace KL divergence to enabling backpropagation of discrete outputs.

This study applies both GAN and VAE for synthesizing individual-

specific activity schedules with a customized structure for high-dimensional discrete data. High-level descriptions of the theory behind the VAE and GAN are described in the following subsections. More detailed discussion and mathematical rigor are presented in Kingma and Welling (2013) for VAEs and Goodfellow et al. (2016) for GANs.

4.3. Generation Module (GM)

4.3.1. Variational autoencoder

The VAE estimates the mapping function of the distribution of the latent variable Z into the data space to approximate true data distribution $P(X)$ (i.e., variational inference). Given some known $P(Z)$, such as multivariate Gaussian, the $P(Z)$ can be mapped using the estimated non-linear mapping function to approximate true data distribution $P(X)$. The VAE generates the synthetic samples by sampling the data from $P(Z)$ and mapping it to the X with $P_\phi(Z)$ (i.e., "decoder"), where ϕ are its estimated parameters. To efficiently estimate the ϕ, alternative function $\mathrm{Q}_\theta(X)$ with parameter θ is applied (i.e., "encoder"). The encoder maps X to Z, while the decoder maps Z to X. During the training stage that estimate θ and ϕ, the encoder maps an input data to the latent space, which in turn is mapped back to the observed data space by the decoder. The difference between the input and output data is calculated as an error for training. This joint estimation of $P_\phi(Z)$ and $\mathrm{Q}_\theta(X)$ makes the model efficiently represent the data in the latent space.

The decoder and encoder could be any non-linear mapping functions according to data space. Since the target SC and NHTS data are high-dimensional discrete data, we use fully connected multilayer artificial neural networks (ANN). The decoder and encoder architecture usually has a bottleneck structure with the dimension of Z being much less than those of X, and the decoder and encoder have the mirror form. Representation of sparse data can be effectively learned in the low-dimensional latent space using this bottleneck structure (Borysov et al., 2019).

Conditional VAE (CVAE) is an extension of VAE that controls the data generation process based on given input observation (Sohn et al., 2015). The input observation modulates as the prior information for Gaussian latent variables that generate outputs. The CVAE is applied to both the GM and LM. In the GM, individual-specific activity schedules are factorized into ISA and ATA to reduce the data dimension. The CVAE estimates the joint probability distribution of NHTS data as the conditional probability of ISA given ATA, or vice versa. The remaining joint probability of ATA or ISA is estimated using VAE or other generative models. Figure 3 represents the model structure of CVAE of which the encoder and decoder are constructed using fully connected ANN.

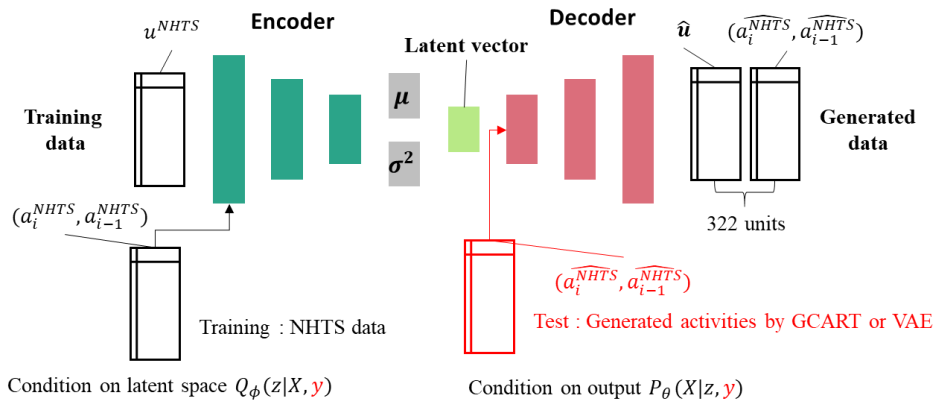


Figure 3. Model structure of the generation module using conditional VAE

4.3.2. Generative adversarial networks

The GAN's basic idea is simultaneous training of the generator and discriminator networks using a min-max two-player game. The generator learns to generate samples from latent space, equivalent to the sample drawn from the same distribution of real samples. In contrast, the discriminator learns to distinguish real samples from the generated ones. The generator and discriminator are represented as a structure of neural networks such as CNN, RNN, and fully connected ANN as in VAE. The solution of the two-player game is a Nash equilibrium where the discriminator cannot discriminate the real samples from the generated samples while the generator generates the realistic samples.

Mathematically, the GAN estimates the generated data distribution P_g by approximating true data distribution $P(X)$ using a parametrized generator $G(z;\theta_g)$ with z drawn from a multivariate random distribution P_z.

A parametrized discriminator $D(\mathrm{X};\theta_d)$ is trained with the following value function as in Equation 1 (i.e., a min-max two-player game)

$$\min_{\theta_d}\max_{\theta_g} E_{x\sim P_{data}}[D(x)] + E_{z\sim P_z}[1 - D\big(G(z)\big)] \tag{1}$$

Training the discriminator until convergence by minimizing the value function is the same as minimizing the Jensen-Shannon divergence (JSD) between P_{data} and P_g. However, it can suffer vanishing gradients problems, which prevent the parameter from changing its value. Therefore, two mini-batches of size m are sampled from the training data and the random distribution. The model is trained by alternating discriminator and generator learning steps with these two mini-batches optimizing a separate loss for each case as in Equation 2.

$$\begin{aligned}\mathcal{L}_{\mathrm{d}} &= \frac{1}{m}\sum_{i=1}^{m}\log\big(D(X_i)\big) + \log(1 - D\big(G(z_i)\big)\,\mathcal{L}_g \\ &= \frac{1}{m}\sum_{i=1}^{m}\log\big(G(Z_i)\big)\end{aligned} \tag{2}$$

For generating discrete data, P_{data} is discrete on Δ_n^T that is a T-dimensional set of simplex, $\Delta_n = \{p \in R^n : p_i \geq 0,\ \sum p_i = 1\}$, but P_g can be continuous over Δ_n^T. This discrepancy causes the JSD between P_{data} and P_g is infinite.

To address this issue, Wasserstein GAN (WGAN) (Arjovsky et al., 2017) that uses the Wasserstein-1 distance (WAD) instead of JSD was proposed since the WAD is continuous everywhere and differentiable almost everywhere. The authors also show that the Wasserstein loss is more correlated with sample quality (i.e., the loss is tractable). In practice, the critic replaces the discriminator, which returns real values about the similarity between the real and fake sample instead of binary outputs (i.e., true or not). The loss function of WGAN is described as in Equation 3.

$$\mathcal{L}_{\mathrm{d}} = \frac{1}{m}\sum_{i=1}^{m} -D(x_i) + D(G(z_i))\, \mathcal{L}_g = \frac{1}{m}\sum_{i=1}^{m} -D(G(z_i)) \tag{3}$$

WGAN-GP (Gulrajani et al., 2017) is modified WGAN, adding gradient penalty to the critic loss, enabling stable GAN training with almost no hyperparameter tuning for discrete data with a continuous generator. This study adopts the WGAN-GP for synthesizing individual-specific activity schedules. Conditional extension of WGAN-GP (CWGAN-GP) is also implemented to evaluate the efficiency of factorization of ISA and ATA. Figure 4 represents the model structure of CWGAN-GP, of which the discriminator (i.e., the critic in WGAN-GP) and generator are constructed using fully connected ANN. Like CVAE, the CWGAN-GP estimated the joint probability distribution of NHTS data as the conditional probability of ISA

given ATA, or vice versa. The remaining joint probability of ATA or ISA is estimated using WGAN-GP or other generative models.

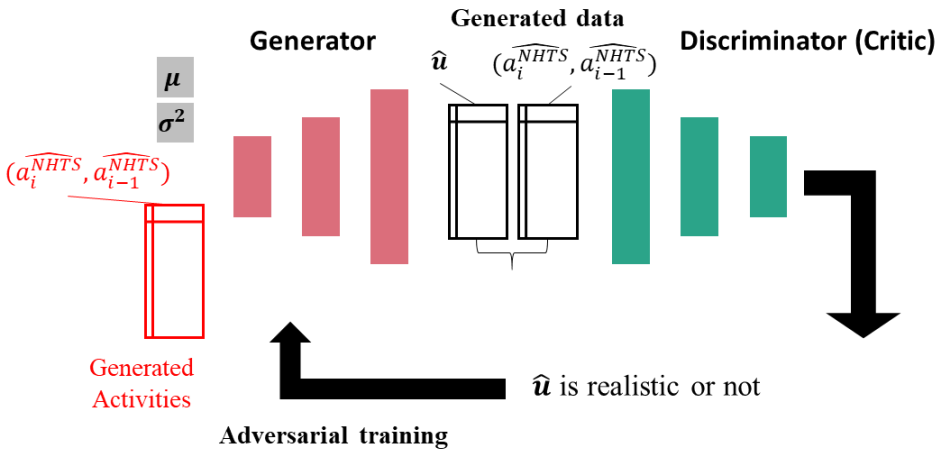


Figure 4. Model structure of the generation module using conditional WGAN-GP

4.3.3. Gibbs sampler with CART

Gibbs sampler works by estimating joint probability distribution based on a series of conditional probability distributions of n variables conditioned on the other variables. One variable y_{obs}^1 of the data is selected, and the distribution of this variable, conditioned on the other variable, $P(y_{obs}^1|x_{obs})$, is estimated. Then, the distribution of the next variable is estimated given other variables and previously estimated variable $P(y_{obs}^2|y_{obs}^1, x_{obs})$. The distribution of the subsequent variable of y_{obs} is estimated conditional on x_{obs} and all previously estimated variables of y_{obs}. The generation of the synthetic data sets proceeds in parallel to the fitting of each conditional distribution. Each variable of the synthetic data is generated from the assumed

distribution, conditional on x_{obs}, the fitted parameters of the conditional distribution, and the synthesized values of all the previously estimated variables of y_{obs}.

This study estimates the conditional distribution using classification and regression tree (CART)(Drechsler and Reiter, 2011; Nowok et al., 2016). The objective of the CART algorithm is to find the subsets of the data space that have relatively homogeneous outcomes. Recursive binary splits of the variables found the subsets and a decision tree with leaves corresponding to these subsets model the series of binary splits. When the CART algorithm is conducted, the values in each subset (i.e., leaf) indicate the conditional probability distribution of the output for the subsets in the data, given input variables that satisfy the criteria that define the leaf.

4.4. Merging Module (MM)

The GM generates the synthetic NHTS in the form of consecutive trips, and these consecutive trips are interconnected to create the synthetic NHTS in the form of a trip-chain. The MM is a trip-chaining algorithm for the generated consecutive trips, which uses the estimated joint probability distribution of the GM. Detail procedure of the MM can be described as follows. First, many synthetic NHTS data are sampled from the joint probability distribution of the GM. The samples are then divided according to the number of trips (*N)* within a day since the travel patterns of trip-chain with $N = 2$ and those with

$N > 2$ could be significantly different. While the former represents the regular commute and returning home trips, the latter can indicate various travel patterns. Figure 5 compares the marginal distribution of activity trip attributes extracted from NHTS data, which support the difference of travel patterns between trip-chain with N = 2 and N > 2. For example, the distribution of age shows that young people tend to travel more than older people. The activity duration is significantly shorter in trip-chain with $N > 2$ than trip-chain with $N = 2$. These differences can be caused by trips to private educational institutes, shopping, and leisure. Also, the first, last, and other activities are sampled from the synthetic NHTS, respectively.

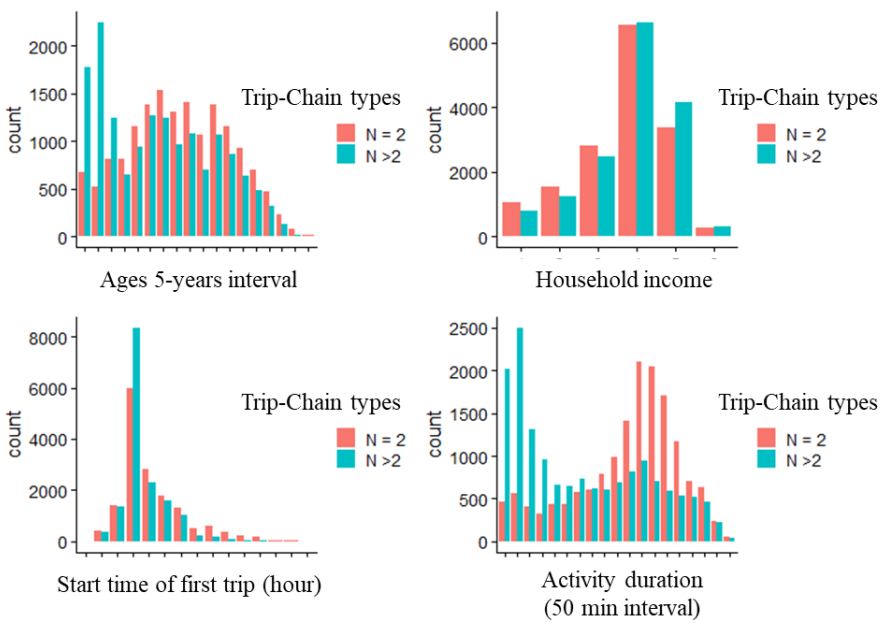


Figure 5. Comparison of marginal distribution of activity-trip attributes extracted from NHTS data: trip-chain with N = 2 and N > 2

Based on the classified samples, two cases of MM are proposed and evaluated,

as depicted in Figure 6. For case 1, the first activity is sampled, and the corresponding next trip is found by matching the present ATA of the first activity and the prior ATA of the next activity. Among the matched next activity, the sample that has the most similar ISA to the first trip is selected as the next activity. The similarity of ISA is calculated by *Eskin* measure (Eskin et al., 2002), which assign higher weights to mismatches by attributes with a higher number of categories. The similarity of two different categorical variables, x_{ic} and x_{jc}, is calculated by Equation 4.

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \;\; if \;\; x_{ic} = x_{jc} \\ \dfrac{n_c^2}{n_c^2 + 2} \;\; \text{otherwise} \end{cases} \tag{4}$$

where n_c is the number of categories of the c-th variable. These processes are repeated *N*-2 times and matching the last trip to the corresponding *N*-1th trip. When the *N* = 2, the last trip is directly matched to the first trip. The connected trip-chain has a different ISA since the trip with the most similar ISA is selected as the next trip. The representative ISA of the trip-chain is selected by sampling each ISA among different ISA in the trip-chain. For example, if the gender of trip-chain with N = 4 are 3 men and 1 woman, sampling one gender and select it as a representative gender. This sampling allows the selection of representative ISAs while maintaining the original probability distribution of ISAs of synthetic NHTS data. Case 2 is the same

as case 1, except the last activity is sampled instead of the first activity, and the remaining process is conducted in the reverse direction.

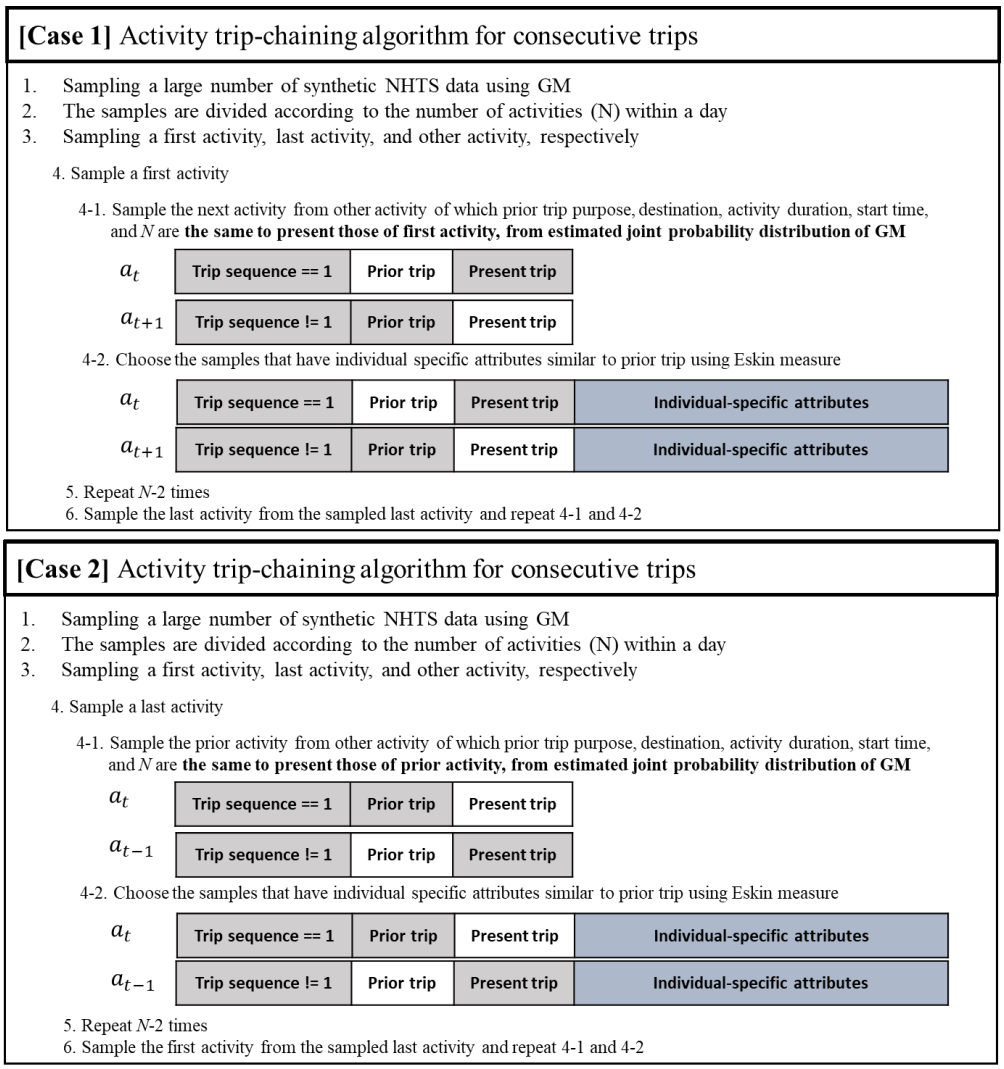


Figure 6. Activity trip-chaining algorithm for consecutive trips based on the joint probability distribution of the GM

4.5. Labeling Module (LM)

The LM estimates the ISA of SC data (i.e., incomplete sample data) given the ATA of SC data. The probability distribution of ISA given ATA, observed in the NHTS data, is estimated by conditional DGM. Among various conditional DGMs, a model that estimates the most realistic ISA in the NHTS data is selected as a model for the LM under the assumption that the relationship between ATA and ISA is equivalent in NHTS and SC data. The performance evaluation of conditional DGMs shows that CVAE is the best model structure for estimating the probability distribution of ISA given ATA. The model structure of the CVAE is presented in Figure 7. The trained decoder uses the random vectors following a multivariate normal distribution and the ATA of NHTS data $\left(a_i^{NHTS}, a_{i-1}^{NHTS}\right)$ or SC data (a_i^{SC}, a_{i-1}^{SC}) as inputs to estimate corresponding ISA ($\hat{\boldsymbol{u}}$).

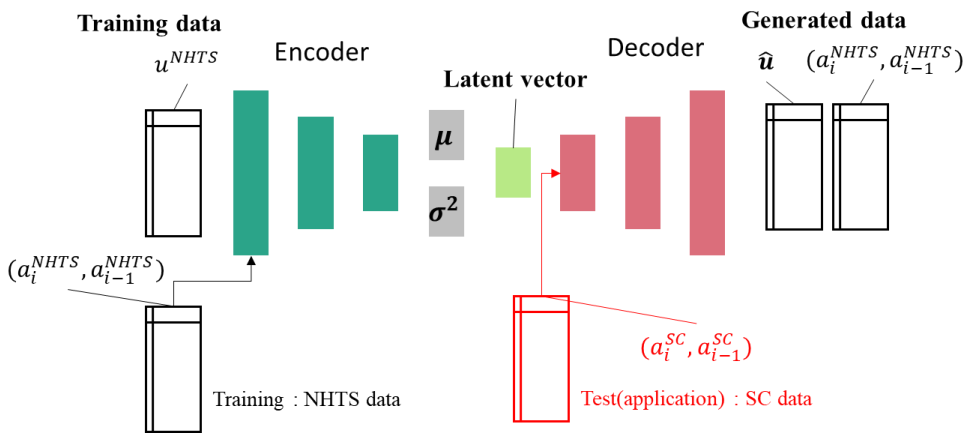


Figure 7. Model structure of the labeling module using conditional VAE

4.6. Evaluation metrics

The GM and LM commonly aim at estimating the probability distribution of the complete sample data; thus, all the performance evaluations in this study are based on measuring the similarity of probability distribution between the original data and estimated data. Specifically, the GM is evaluated from the perspective of the probability distribution and the applicability to ABM. The probability distribution refers to measuring the similarity of marginal and bivariate distribution of each categorical attribute. The applicability to ABM indicates the similarity of a probability distribution of specific attributes for ABM. The LM is evaluated in terms of the probability distribution and the applicability to travel behavior analysis. The latter indicates the rationality of travel behavior analyzed by the generated ISC from LM. This measure is evaluated based on expert judgment and qualitative comparison with other behavior analysis.

Quantitative evaluation of similarity of the probability distribution is conducted using Jenson-Shannon divergence (JSD). The JSD measures how well the statistical properties of the real data are captured and transferred to synthetic data. The JSD is based on KL divergence but is symmetric and finite value ranged from zero to one, which can be used as an evaluation metric. Equations 5 and 6 describe how the KLD and JSD are calculated in the categorical data, where $|v|$ is the number of levels of categorical variables, and P_v and Q_v are probability mass functions of real and generated data,

respectively.

$$D_{KL}(P_v||Q_v) = \sum_{i=1}^{|v|} P_v(i) log \frac{P_v(i)}{Q_v(i)} \quad (5)$$

$$D_{JS}\,(P_v||Q_v) = 0.5[D_{KL}(P_v||0.5(P_v + Q_v)) + D_{KL}(Q_v||0.5(P_v + Q_v)) \quad (6)$$

The JSD only can be defined at the variable level, not over the entire dataset. Therefore, to evaluate the overall similarity of different categorical attributes, the averaged JSD of all categorical variables is used to evaluate generation performance. Also, I measure the similarity of the marginal distribution of categorical variables and the bivariate distribution of them to evaluate how well correlations among the variables were able to capture. All combinations of categorical variables are considered in the evaluation (e.g., $n\mathrm{C}_2$ combinations for n categorical variables).

The applicability to ABM is an important measure since it can directly evaluate the usability of the generated data. Synthetic NHTS data in the form of trip-chain are evaluated from this perspective, which indicates the performance of both GM and LM. I modified the validation framework for an activity-based model (VALFRAM)(Drchal et al., 2016) to fit the context of this study. VALFRAM is a set of steps comparing the probability distribution of selected attributes of the activity chain. Table 5 represents the nine selected attributes of VALFRAM consisted of six trip attributes and three trip-chain attributes. Trip-chain attributes only can be extracted from trip-chain data,

which directly measures the performance of MM. Each attribute is defined by marginal, bivariate, trivariate, or quadvariate distribution of the categorical attributes, and it is noted in the fourth column of Table 5. The similarity of those attributes between original and synthetic data are evaluated using JSD.

Table 5 Modified VALFRAM for evaluating synthetic NHTS data in the form of trip-chain

Type	Attributes	Resolution	Dimension	Index
Trip	Trip count for each travel mode	Time of day (1 hour)	Bivariate	T-1
		361 grids	Trivariate	T-2
	Trip count for each purpose	Time of day (1 hour)	Bivariate	T-3
		361 grids	Trivariate	T-4
	OD Matrix	361 grids by 361 grids	Quadvariate	T-5
	Activity duration	Trip purpose (3)	Bivariate	T-6
Trip-chain	Activity sequence	Individual	Marginal	TC-1
	Total travel time	Individual	Marginal	TC-2
	Total activity duration	Individual	Marginal	TC-3

Chapter 5. Experimental Results

5.1. Generation Module (GM)

5.1.1 Hyper-parameter tuning

Tuning hyper-parameter is important for the reliable and generalizable performance of the deep learning model. It should be conducted in a systematic and reproducible way to prevent under or overestimating the performance of the benchmark models. For the GCART, a default parameter settings for CART suggested in Nowok et al. (2016) is applied, and the sequence of the conditional probability of Gibb sampler is calibrated. After several experiments with GCART, I have derived that $P(ASA_t|ASA_{t-1}|ISA)$ is the best sequence for the GM. The sequence of each attribute within ASA or ISA did not significantly change the performance. For the two DGM, VAE and WGAN-GP, I calibrated the structure of ANN (i.e., the number of hidden layers and neurons) making up the encoder and decoder (in case of VAE), or the generator and critic (in case of WGAN-GP). Other hyper-parameters such as optimizer, learning rate, and batch size are set as a recommended value in the previous studies (i.e., Gulrajani et al. (2017) for WGAN-GP and Kingma and Welling (2013) for VAE).

To optimize the network structure of the VAE, grid search is conducted for the fully connected ANN architecture consisting of 1 to 3 hidden layers with 10, 25, 50, 100, or 200 neurons, respectively. The

bottleneck structure is used for the encoder and decoder, of which the number of neurons in the hidden layer of the encoder decreases in the next hidden layer. Also, the decoder is a mirror of the encoder. For example, if the encoder has three hidden layers with 50, 25, and 10 neurons, the decoder has three hidden layers with 10, 25, and 100 neurons. Grid search is also applied to optimize the network structure of the WGAN-GP. The grid search considers the fully-connected ANN consisting of 1 to 3 hidden layers with 64, 128, 256, and 512 neurons.

5.1.2. Generated sample and population

One of the main hypotheses in this study is that the synthetic NHTS data sampled from a joint probability distribution of NHTS would represent the population better than the original NHTS data since the generated data can represent the missing activities from original NHTS data but exist in the population data. Verifying this hypothesis requires the population data of NHTS, but those data do not exist. Alternatively, I used the 100% SC data, which represent the population of transit travel. Using these population data, the samples from different strategies are generated as follows: (i) SC data sampled from 1% to 10%, (ii) original NHTS data only for transit travel, and (iii) synthetic NHTS data only for transit travel. The synthetic NHTS data are generated from the estimated probability distribution of NHTS using VAE. The best-performed structure of VAE for estimating the joint probability distribution of transit NHTS data are three hidden layers with 100, 50, and 25.

One day SC data on Nov 16, 2016, in Seoul were used as population data for transit travel. The data include 15 categorical variables (i.e., ATA of SC data) with 174 dimensions. The average JSD of 15 categorical variables is used to evaluate the differences in statistical distribution between the population and sample data. Each sampling strategy is repeatedly conducted ten times with replacement (i.e., bootstrapping) to estimate the statistical significance of the average JSD. Table 6 shows the average JSD of marginal and bivariate distribution between the population and each sampling strategy. In the case of sampled SC data, as expected, the difference with the population is reduced proportionally to the sample rate. While the average JSD of 1% SC data is 2.80e-05, 10% SC data is 2.59e-06.

In NHTS data, the statistical similarity of NHTS data to the population is much lower than those of sampled SC data. This result indicates that the expected bias of NHTS data, such as response bias participants and cognitive errors, have a significant impact on the behavioral attributes. The generated NHTS data represent the SC better than the original NHTS data in terms of average JSD of marginal and bivariate distribution. The improvements of the generated NHTS are statistically significant at the 1% level. This result verifies the hypothesis of which the generated NHTS data improve the quality of NHTS data. In other words, the statistical bias of the travel survey in representing revealed preference can be mitigated by data synthesis.

Table 6 Average JSD of marginal and bivariate distribution between the population and each sampling strategy

Attributes	Average JSD (Marginal)		Average JSD (Bivariate)	
	Mean	Std	Mean	Std
SC data (1%)	2.80e-05	3.05e-06	2.87e-04	8.22e-06
SC data (5%)	5.39e-06	6.70e-07	5.70e-05	1.53e-06
SC data (10%)	2.59e-06	2.87e-07	2.67e-05	7.09e-07
NHTS data (100%)	1.62e-02	-	3.56e-02	-
Generated NHTS data from VAE	1.53e-02	1.46e-04	3.38e-02	2.69e-04

※ Difference of average JSD (Marginal and Bivariate) between NHTS data and Generated NHTS data is statistically significant at 1% level.

Figure 8 presents how such improvement of the generated data can be achieved by comparing the marginal distribution of SC data, generated NHTS, and original NHTS data. Visual inspection of these distributions reveals that improved quality of generated NHTS may be due to the smoothed properties of generated data. By estimating probability distribution, some of the prominent behavior has been modified to rational behavior, making the generated data more similar to the population.

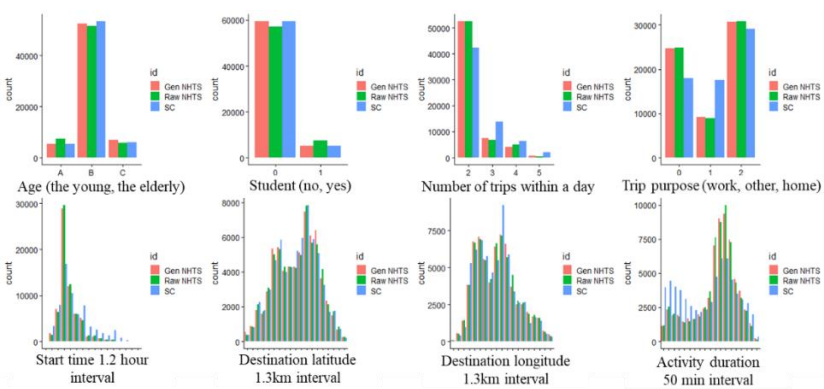


Figure 8. Marginal distributions of the SC data, generated NHTS data, and original NHTS data

5.1.3. Synthetic NHTS data (consecutive trips)

The objective of GM is estimating the full joint probability distribution of NHTS data that includes 28 categorical variables with 322 dimensions (i.e., 322 one-hot encoded variables). I selected three state-of-the-art generative models as benchmark models, i.e., GCART, VAE, and WGAN-GP. I also considered the hybrid model consisted of the ATA (or ISA) generator and the ISA (or ATA) generator using given ATA (or ISA). The former's models were GCART, VAE, and WGAN-GP, while the latter's models were CVAE and CWGAN-GP. Various network structures (i.e., hyper-parameter) of GAN and VAE were evaluated using average JSD for marginal and bivariate distributions. The best-performed VAE had three hidden layers with 200, 100, and 50, and 25 latent dimensions. The CVAE for the hybrid model, which estimates the probability of ATA given ISA or ISA given ATA, also had three hidden layers with 200, 100, and 50, and 25 latent dimensions. While the best-performed WGAN-GP had three hidden layers with 256 neurons (i.e., layers(256,256,256)), the CWGAN-GP for the hybrid model, which estimates the probability of ATA given ISA or the probability of ISA given ATA, had two hidden layers with 256 neurons. The CVAE is additionally applied to the hybrid model, which estimates the probability distribution of ATA_t given ATA_{t-1}, and ATA_{t-1} given ISA. The model structure of those cases were both three hidden layers with 100, 50, 25, and 10 latent dimensions.

Table 7 shows the evaluation results for generating synthetic NHTS data in the form of consecutive trips. As a single model, GCART

outperformed the VAE and WGAN-GP in estimating both marginal and bivariate distribution. As the GCART assumes, the result indicates that a set of conditional probability distributions would be an efficient approximation for the joint probability distribution of the NHTS data. The hybrid model achieved the best performance of GM. The model that generated the ATA using GCART and generated the ISA using CVAE shows the best performance among the various benchmark models. The GCART also outperformed the VAE in the tasks of generating ATA or ISA, respectively. The fact that marginal JSD for ATA of GCART for full data (4.59e-04) is higher than those of GCART for ATA data (2.35e-04) indicates that the performance of GCART decreases for the high-dimensional problem. Therefore, the CVAE complements the GCART by factorizing ISA and ATA, which effectively improves performance. Among the hybrid models of CVAE, the model estimating the probability of ISA given ATA shows better performance than those estimating the probability of ATA given ISA. Note that this best-performed structure of the CVAE was the same model structure of the LM. This result implies that the proposed LM is not only practically useful but also a reasonable structure for estimating the probability distribution of ISA given ATA.

Table 7 Average JSD of marginal and bivariate distribution in synthesizing NHTS data of consecutive trips

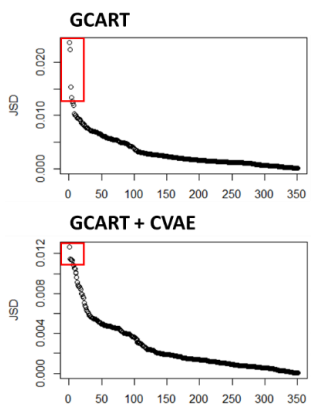
Sample data	JSD (Marginal)			JSD (Bivariate)
	ISA (Mean)	ATA (Mean)	Sum (Mean)	Sum (Mean)
Training data	5.95e-05	2.79e-04	1.81e-04	1.29e-03
Gibbs using CART (GCART)	5.87e-04	4.59e-04	5.16e-04	3.23e-03
Full VAE	1.92e-03	1.63e-03	1.76e-03	8.45e-03
CVAE (ISA | ATA)×GCART (ATA)*	8.17e-04	2.35e-04	4.94e-04	2.72e-03
CVAE (ISA | ATA)×VAE(ATA)	8.71e-02	1.62e-03	3.96e-02	1.02e-01
CVAE (ATA | ISA)×GCART (ISA)	2.70e-04	1.24e-03	8.14e-04	3.63e-03
CVAE (ATA | ISA)×VAE(ISA)	2.19e-03	2.69e-03	2.47e-03	9.48e-03
CVAE (ATAt | ATAt-1)×CVAE (ATAt-1 | ISA) ×GCART(ISA)	2.70e-04	2.15e-03	1.31e-03	6.81e-03
Full WGAN-GP	1.54e-02	1.66e-02	1.60e-02	4.01e-02
CWGAN-GP(ISA | ATA)× GCART (ATA)	8.79e-03	2.35e-04	4.04e-03	1.07e-02

Although the advanced structure of GAN (i.e., WGNGP and CWGAN-GP) are applied, the CVAE and VAE, which are known to stable structure for discrete data (Borysov et al., 2019), outperformed the CWGAN-GP and WGAN-GP. GAN's inferior performances for discrete data were also reported in the other state-of-the-art GAN models (Xu et al., 2019). Therefore, future improvements in the model will require a review of the model structure for the discrete data based on the VAE.

A more fragmented form of the hybrid model was considered in the experiments. The consecutive activity trips were factorized into each activity trip, and the hybrid model consisted of CVAE for $P(ATA_t|ATA_{t-1})$, CVAE for $P(ATA_{t-1}|ISA)$, and GCART for $P(ISA)$ was evaluated. The

performance of that fragmented hybrid model shows lower performance than the hybrid model consisted of CVAE for $P(ATA|ISA)$ and GCART for $P(ISA)$. This result reveals that the factorization for simplifying model structure does not always involve performance improvement, which may be since the activity schedule within a day was not determined in the sequence of activity trip. In other words, individuals are more likely to schedule based on priorities than in a purely sequential fashion (Doherty, 2000).

Figure 9 visualizes the distribution of bivariate JSD of GCART and hybrid model of GCART and CVAE to analyze the performance difference. It shows that the performance difference between those two models comes from destination choices, which are the most difficult part in ABM (Rasouli and Timmerman, 2014). Therefore, additional techniques to address the destination choice problem need to be considered to improve the model further.



Rank	Variable 1	Variable 2	JSD
1	Destination latitude (*t*)	Destination longitude (*t*)	0.024
2	Destination latitude (*t-1*)	Destination longitude (*t-1*)	0.022
3	Household income	Start time	0.015
4	Home longitude (*t*)	Work latitude (*t*)	0.013
5	Work latitude (*t*)	Work longitude (*t*)	0.013

Rank	Variable 1	Variable 2	JSD
1	Age	Student	0.013
2	Home latitude (*t*)	Home longitude (*t*)	0.013
3	Work latitude (*t*)	Work longitude (*t*)	0.012
4	Home longitude (*t*)	Work longitude (*t*)	0.012
5	Home latitude (*t*)	Work longitude (*t*)	0.012

Figure 9. Distribution of bivariate JSD of GCART and hybrid model of GCART and CVAE

Figure 10 depicts the comparison of marginal and bivariate (or trivariate) distribution between test NHTS data and generated NHTS data. Parentheses indicate the JSD of each variable. The marginal distribution of the generated NHTS was very similar to those of the test data, but trivariate distribution (i.e., distribution of destination of each trip purpose) shows quite different patterns. However, those differences may not be significant from a macroscopic perspective in transportation planning. To evaluate the estimated probability distribution from the perspective of the applicability to ABM, the synthetic NHTS data were processed by the MM to complete the activity-chain, and they were evaluated using VALFRAM. These contents are provided in the next subsection.

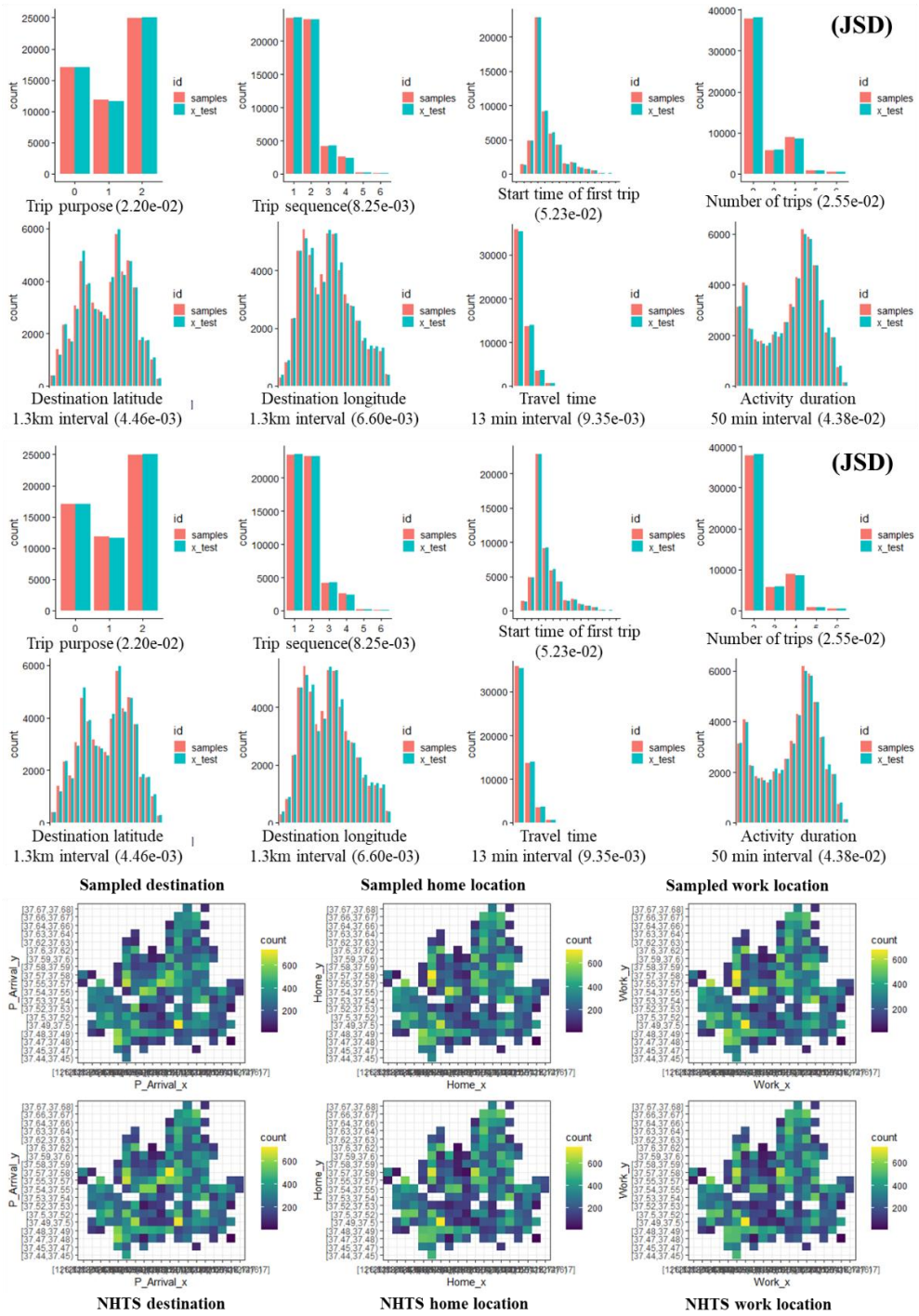


Figure 10. Comparison of marginal, bivariate, and trivariate distribution between test NHTS data and generated NHTS data

5.2. Merging Module (MM)

The GM generates the synthetic NHTS data in the form of consecutive activity trips, and the hybrid models combining GCART and VAE show the best performance in synthesizing those data. The MM merged the synthetic consecutive trips to create the synthetic activity trip-chain. Two cases of the MM divided by trip-chain order (i.e., from the first trip or from the last trip to construct the trip-chain) were considered in this study. The created synthetic activity trip chains are evaluated using JSD for selected attributes in VALFRAM to investigate the applicability to ABM.

Using GM and MM, I created the synthetic NHTS data of 48,824 trips. Table 8 shows the evaluation results of VALFRAM for six trip-related attributes and three trip-chain-related attributes. Overall, the MM of case 1 outperformed the MM of case 2, which indicate that chaining by trip-sequence is more reasonable than those by reverse trip-sequence. Note that trip-chaining can be conducted by considering the trip-chain's overall context, but deriving those contexts requires long-term continuous trip-chain data. If the MM is applied to other data sources, not the NHTS that collects the trip-chain of a single day, other algorithms can be performed.

Generally, the VALFRAM attributes of synthetic data extracted from the marginal probability distribution are expected to more similar to those of original data than the attributes from bivariate or trivariate distribution. However, the attributes from bivariate and trivariate distributions (i.e., T-1, T-

3, and T-6) show comparable similarity to the attributes from marginal distribution. It may indicate that GM and MM can successfully estimate the bivariate and trivariate probability distribution. But the similarity drastically decreases for quadvariate distribution, which requires extracting the origin and destination of trips. However, those difference was not that large as shown in the parenthesis in Table 7, as 3.15 trips for MM-case 1.

Table 8 Evaluation results of VALFRAM for synthetic NHTS data in the form of activity trip-chain

Type	Attribute index	Training	MM-Case 1	MM-Case2	Dimension
Trip	T-1	6.44e-04	1.54e-02	2.96e-02	Bivariate
	T-2	1.11e-02	2.28e-02	4.43e-02	Trivariate
	T-3	4.32e-04	1.50e-02	6.67e-02	Bivariate
	T-4	6.45e-03	1.33e-02	4.95e-02	Trivariate
	T-5	1.48e-01 (2.47)	1.82e-01 (3.15)	2.28e-01 (4.31)	Quadvariate
	T-6	3.84e-04	5.53e-03	5.14e-02	Bivariate
Trip-chain	TC-1	4.12e-04	3.30e-03	6.69e-02	Marginal
	TC-2	2.11e-04	5.13e-03	2.31e-02	Marginal
	TC-3	2.60e-04	1.03e-02	2.57e-02	Marginal

Note : parenthesis indicate the RMSE of the number of OD trips

Since the physical meaning of average JSD is not clear, I visualize the marginal, bivariate, and trivariate distribution captured by attributes in VALFRAM, as shown in Figure 11. For bivariate distributions of trip-related attributes presented in T-1, T-3, and T-6, the visual inspection of the distribution shows that the test and synthetic data have similar distributions even if the value of average JSD were high than 1.54e-02. Even in the trivariate distributions of trip-related attributes (T-4), their estimated travel

patterns still share similar properties. In specific, estimating T-4, which is an activity-specific number of trips, is an important task for transportation planning since it is used for managing peak-hour congestion (commute trips) and analyzing commercial area (other trips). The attributes from marginal distribution are well estimated, as expected, though they are trip-chain-related attributes. This result implies that the proposed MM works well for the synthetic NHTS data of consecutive trips.

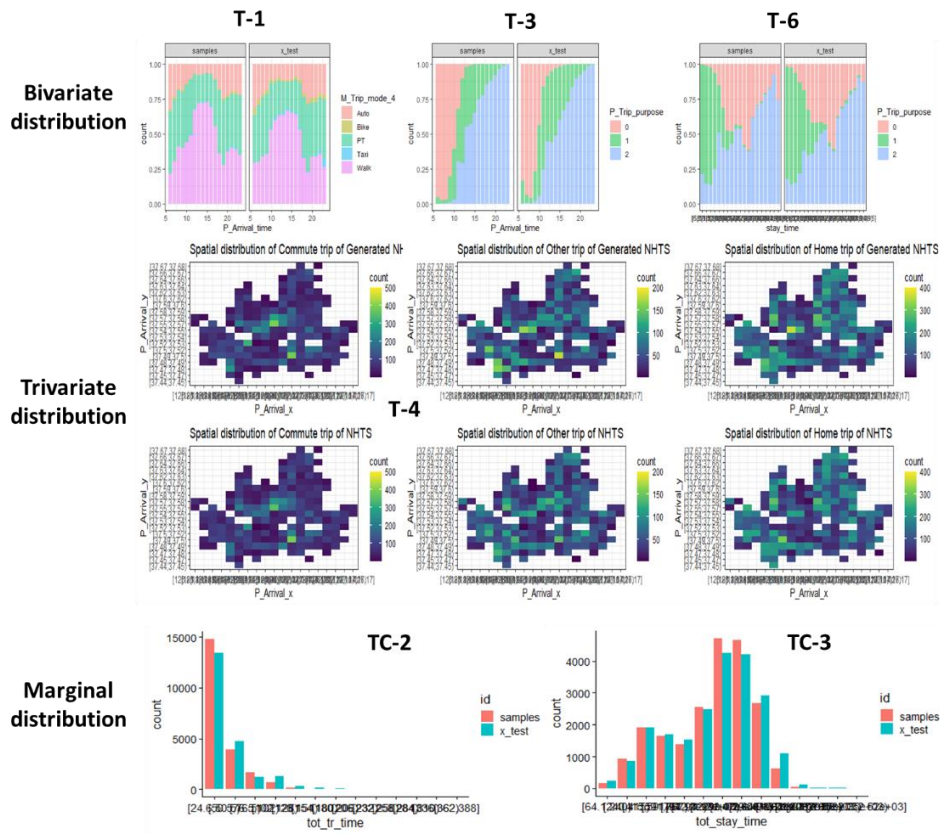


Figure 11. Marginal, bivariate, and trivariate distribution captured by attributes in VALFRAM

The daily activity sequences provide explicit details on activity type, frequency, and sequence of activities, which are core components in any ABM, as their travel demand is determined by those sequences of activity (Hafezi et al., 2018). Table 9 compares the distribution of activity sequences extracted from original NHTS data and synthetic NHTS data. Although activity sequence can be captured from marginal distribution, it is a direct indication of the performance of MM because it contains sequencing information. To remove the outlier patterns, the activity trip chains that have more than 5 trips are excluded. The extracted sequences of activity from synthetic NHTS data show that the MM properly preserves the sequence information of consecutive trips, although it merges each trip in the temporal order. This result can directly show the excellence of the MM.

Table 9 Distribution of activity sequence extracted from original NHTS data and synthetic NHTS data.

Activity sequence	Original NHTS	Synthetic NHTS from GM and MM
0-2	59.3	61.0
1-2	23.8	21.7
0-1-2	6.2	5.7
0-2-1-2	4.3	5.9
0-1-1-2	2.3	1.6
1-1-2	2.1	2.6
1-2-1-2	1.8	1.1
1-1-1-2	0.2	0.3
1-2-0-2	0.1	0.0

Note : 0: Commute, 1: Other, 2: Returning home

5.3. Labeling Module (LM)

The LM estimates the probability distribution of ISA given ATA of SC data. Similar to the GM, the performance of the LM is estimated using JSD of marginal and bivariate distributions between ISA of NHTS data and estimated ISA of SC data. Since these two data are obtained from different sources, the population distribution of these two data could be fundamentally different. For example, while the SC data records all trips, including minor ones, the NHTS can omit some of the minor trips by respondents. In addition, the record time of NHTS can be distorted by respondents' memory from 30 minutes to one hour. Figure 12 represents these discrepancies by comparing ATA of SC data and ATA of NHTS data, which are given information of the LM. While the distribution of destination and trip sequence are relatively similar, the distribution of start time, activity duration, number of trips, and travel time are quite different due to the aforementioned response bias from the travel survey. To control the effect of this difference in population distribution, I resample the start time and activity duration of SC data to match the distribution of those attributes to the distribution of NHTS one.

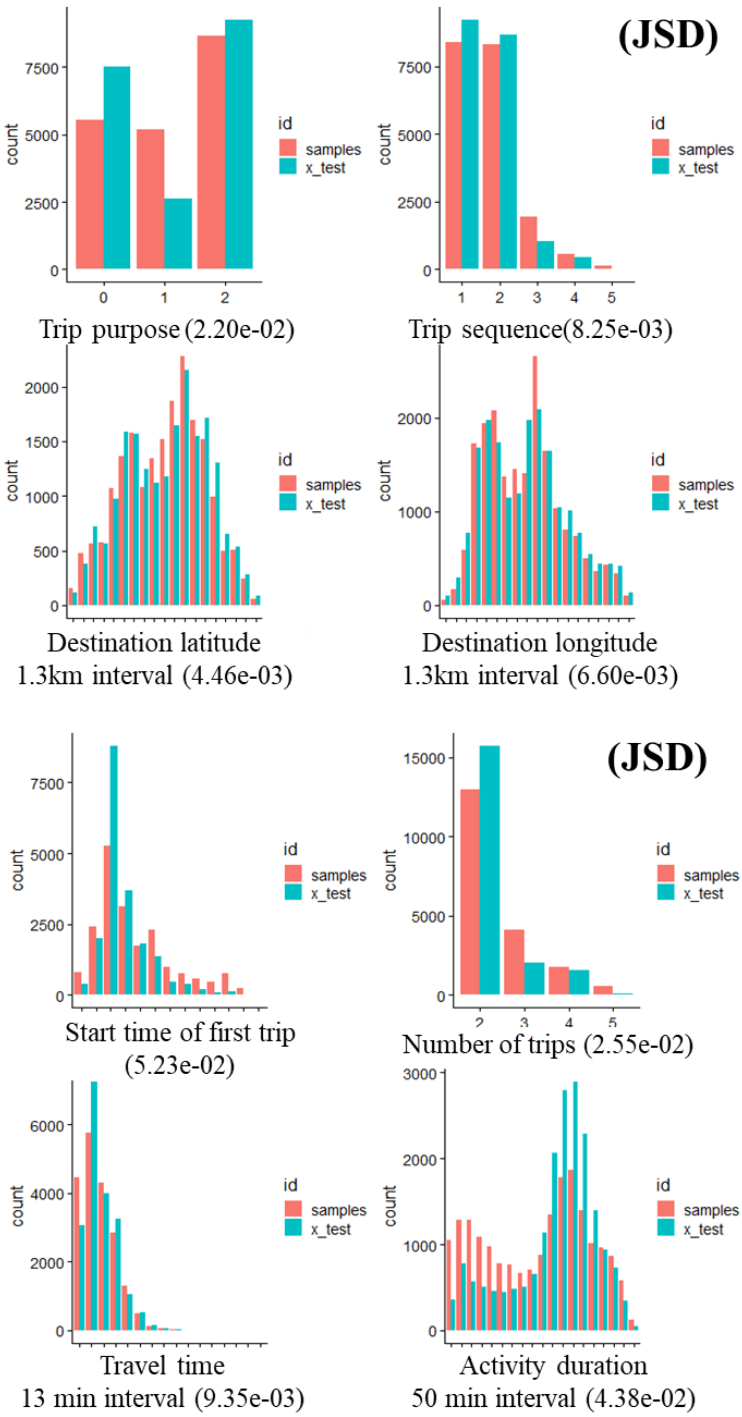


Figure 12. Differences of marginal distribution between SC and NHTS data.

Table 10 shows the performance of estimating ISA of SC data given ATA. Estimated ISA of SC shows comparable results (3.72e-03) to those of NHTS (1.01e-03) considering the difference of ATA between SC and NHTS (1.82e-02). The CVAE outperformed the CWGAN-GP, as in the GM, for estimating the probability distribution of ISA given ATA. It indicates that the strength of the VAE for discrete data over WGAN-GP is also valid to the labeling task. I resampled the SC data to control the population differences between SC and NHTS. Multivariate control for resampling is a non-trivial task since it requires a much larger number of data points. Therefore, the SC data are resampled only considering the NHTS's distribution of the start-time and activity duration. Controlling the ATA reduced the difference of ATA from 1.82e-02 of SC data to 4.22e-03 of resampled SC data; however, such control did not reduce the differences of ISA to NHTS data, rather slightly increases the differences from 3.72e-03 to 4.22e-03. This result may imply that the estimated ATA was not determined by a few attributes due to the complex interrelations of ATA and ISA.

Table 10 Performance of estimating ISA of SC data given ATA

Model	Given attributes (ATA)	Target attributes (ISA)	JSD(Marginal)			JSD. (Bivariate)
			ISA (Mean)	ATA (Mean)	Sum (Mean)	Sum (Mean)
CVAE	NHTS	NHTS	1.01e-03	0	4.82e-04	2.16e-03
	SC	NHTS	3.72e-03	1.82e-02	1.13e-02	2.81e-02
	SC(Resampled)	NHTS	4.22e-03	4.48e-03	4.35e-03	1.49e-02
CWGAN-GP	NHTS	NHTS	4.11e-03	0	1.97e-03	5.15e-03
	SC	NHTS	6.67e-03	1.82e-02	1.27e-02	3.10e-02
	SC(Resampled)	NHTS	6.99e-03	7.99e-03	7.51e-03	2.09e-02

Figure 13 represents the marginal distribution of estimated ISA of SC data, compared with those of NHTS data. Despite the difference in the underlying distribution of ATA of two data, the estimated ISA of SC shows reasonable distribution compared with those of NHTS data. The low dimensional categorical variables show a marginal distribution close to NHTS, and the high dimensional numerical variables also show a reasonable difference to NHTS.

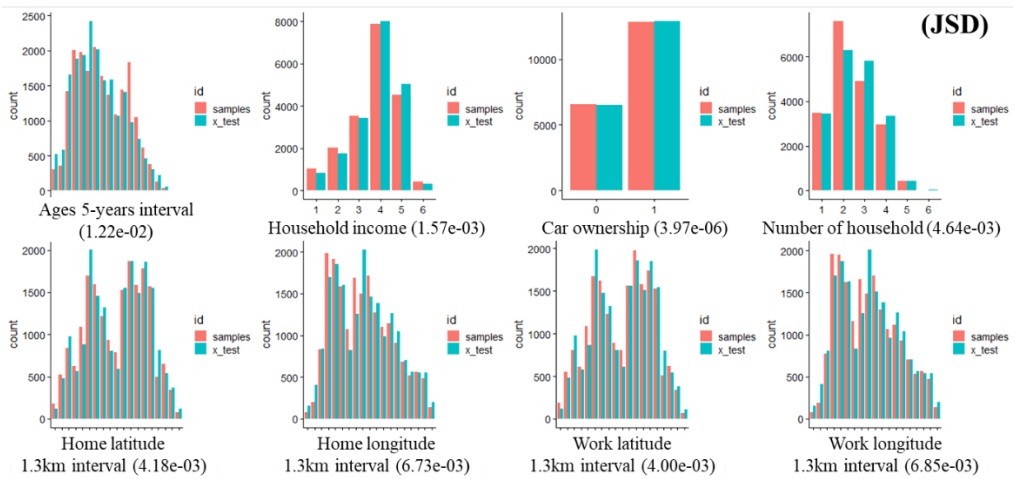


Figure 13. Marginal distribution of estimated ISA of SC data compared with those of NHTS data.

To verify the quality of estimated ISA in general, quantitative validation on population data is required. Some of the ISA, such as age, household income, home types, and driver's license, were collected from Census data, which represent the population of ISC for all people in Seoul. Since the census data were collected by district "Gu," I compare the ISC,

transit NHTS, and Census data according to the "Gu" district. Note that while ISC and transit NHTS include only regular transit users (i.e., people who use transit more than 2 times within a day and more than 3 days during a month.), the census data include every people. Table 11 represents the marginal distribution of estimated household income of SC data, compared with those of NHTS data. While NHTS and ISC showed very high similarities for the overall marginal distribution of household income, those are slightly different for each district. For Seocho-gu, which has a much higher household-income than other districts, the ISC estimated a somewhat lower income than those of NHTS, given the median income from Census. This result may indicate that, like Synthetic NHTS data from GM, the estimated ISC also represented the smoothed properties in the probability distribution.

Table 11 Comparison of household income from ISC, transit NHTS, and Census data

Attribute	Category	Total		Gwanakgu		Seochogu		Nowongu	
		ISC (%)	NHTS (%)	ISC (%)	NHTS (%)	ISC (%)	NHTS (%)	ISC (%)	NHTS (%)
Monthly household income	0	3.9	4.5	3.8	3.3	2.9	1.1	7.7	9.9
	1	9.3	9.1	9.2	8.8	8.8	6.2	10.2	8.5
	2	17.7	18.1	21.0	18.1	13.9	12.2	17.1	20.9
	3	40.7	40.4	42.4	40.8	34.8	28.8	43.5	41.6
	4	26.1	26.1	21.5	27.4	35.2	45.2	19.9	18.7
	5	2.3	1.8	2.2	1.7	4.4	1.7	1.6	0.5
Census median		4.01 million won		3.85 million won		5.10 million won		4.04 million won	
Similarity	JSD	4.03e-04		3.85e-03		1.29e-02		5.40e-03	

Note : 0: less than 1 million won, 1: 100~200 million won, 2: 200~300 million won, 3: 300~500 million won, 4: 500~1000 million won, 5: more than 10 million won.

Table 12 compares the marginal distribution of a driver's license, car availability, and types of house. The distribution of car availability of NHTS and ISC are more similar to those of driver license. The home-type distribution also shows that estimated ISC data represent smoothed but reasonable patterns considering those distributions of Census and NHTS. For example, in Gwanakgu and Nowongu, where the proportion of apartments is significantly higher and lower than the total proportion, respectively, the smoothed properties of ISC estimate a more similar distribution to Census than the distribution of NHTS. These results may suggest that the ISC can provide a better quality of data than NHTS, not just reflect the trend of continuously collected SC data.

Table 12 Comparison of driver license, car availability, and home type from ISC, transit NHTS, and Census data

Attribute	Category	Total			Gwanakgu			Nowongu		
		ISC (%)	NHTS (%)	Cens (%)	ISC (%)	NHTS (%)	Cens (%)	ISC (%)	NHTS (%)	Cens (%)
Driver's license	Yes	35.6	41		35.1	41.3		34.4	40.6	
	No	64.4	59		64.9	58.7		65.6	59.4	
Similarity	JSD	2.22e-03			2.93e-03			2.96e-03		
Car availability	Yes	30.1	33		35	33.8		32.9	32.3	
	No	69.9	67		65	66.2		67.1	67.7	
Similarity	JSD	7.01e-04			1.15e-04			2.95e-05		
House Type	APT	51.7	52.2	42.0	35.9	27.1	23.8	73.5	89.9	76.1
	Villa	15.3	15.0	50.4	19.4	19.9	54.2	7.6	3.8	19.4
	Multi	13.3	12.7		18.3	19.8		6.2	2.3	
	Single	17.2	18.2		23.5	30.3		11.2	2.6	
	Officetels	2.0	1.5	7.7	2.3	2.5	11.9	1.1	1.1	4.6
	Others	0.5	0.4		0.7	0.4		0.4	0.2	
Similarity	JSD	4.69e-04			8.12e-03			3.83e-02		

Note : 0 : less than 1 million won, 1: 100~200 million won, 2: 200~300 million won, 3: 300~500 million won, 4: 500~1000 million won, 5: more than 10 million won.

5.4. Application of Labeling Module

The main outputs of the proposed framework are two kinds of individual-specific activity schedules: (i) synthetic NHTS data generated using GM and MM and (ii) Individual-specific SC (ISC) data generated using LM. This section explores the application of the latter outputs, ISC, which are original data of this study, combining the strengths of NHTS and SC data. The key improvements of the ISC data compared to the NHTS data are a representation of long-term (i.e., monthly or biweekly) travel behavior and the capability of monitoring recent travel behaviors. Based on those improvements, I suggest two applications of ISC: (i) investigation of monthly travel behavior and (ii) investigation of customized travel behavior. Those two applications can provide new insights for travel behavior in urban transportation, as well as replace the costly travel survey for investigating specific travel behavior.

5.4.1. Investigation of monthly travel behavior

In the near future, MaaS offers travelers door-to-door mobility by the combination of different transport modes based on transit. One of the promising aspects of MaaS is subscription planning that reduces the travel cost of combinations of travel modes. The subscription model generally assumes that users pay a monthly fee and receive bundled transit services. MaaS studies rely on the SP data to design subscription plan (i.e., the price for options) (Caiati et al., 2020; Matyas and Kamargianni, 2018; Ho et al.,

2018) because the long-term travel behavior of transit users cannot be obtained by revealed preference data such as SC data that have an anonymous ID without individual attributes. The ISC that can represent transit travel patterns with corresponding individual attributes can directly tackle those issues. The proposed LM applied to one-month SC data representing the monthly travel behavior of transit users in Seoul. The SC data are collected from Nov 1 to Nov 28, 2016. The SC data are filtered out to reveal the expected transit users for MaaS following three steps: (i) Daily trip-chain that have more than one trips and no outliers for start-time and activity duration; (ii) People who traveled within Seoul (i.e., trips starting and finishing at Seoul); (iii) People who traveled more than 3 days during a month. After filtering out the SC data using those three steps, 4,618,591 passenger IDs have remained among 25,113,319 passenger IDs in the Seoul metropolitan area. I sampled 10% of passenger IDs (i.e., 461,859 passengers) to reduce the computational cost of the following analysis.

First, monthly travel behavior in terms of the number of trips and the number of travel days are investigated to compare the statistics of SC from different durations of data. Tables 13 and 14 show several practical implications for designing a subscription plan of MaaS. The number of passenger IDs in one day, one week, two week, and one-month data show that single-day travel data only represent the travel behavior of 47% of the population. In other words, a travel survey collecting one or two days of travel diary can provide only limited information about populations' behavior. The

number of trips within a day can be biased due to the data collection period. Bi-weekly and weekly data overestimate the number of trips within a day, 4% and 15%, respectively. The number of travel days within a week also shows different patterns between weekly and monthly data. Bi-weekly and weekly data overestimate the number of travel days within a month, 42% and 57%, respectively. Above results commonly suggest that pricing for monthly and biweekly plan need to consider the overestimation of the existing single-day data.

Table 13 Comparison of the number of passengers collected from one-day, one-week, bi-week, and one month SC data

Duration	Number of passenger IDs	Ratio to the one-month data
1 day	216,202	47%
1 week	368,854	80%
2 weeks	450,842	98%
1 month	461,859	100%

Table 14 Comparison of the number of trips within a day, the number of travel days within a month collected from one-day, one-week, bi-week, and one month SC data

Duration	Average number of daily trips		The average number of travel days during a week	
	Trips	Ratio to One month	Trips	Ratio to One month
1 day	2.27	236%		
1 week	1.10	105%	3.36	157%
2 weeks	1.00	104%	3.03	142%
1 month	0.96	100%	2.13	100%

Figure 14 shows the distribution of the number of trips and the number of travel days within a month. Monthly travel days and number of

travels have a bimodal distribution. The Gaussian mixture model is applied to classify two groups using the number of trips and the number of travel days. Group 1 in Figure 14 has the number of trips with a mean of 13.8 and standard deviation of 4.73 and the number of days with a mean of 6.2 and standard deviation of 2.12, while Group 2 has the number of trips with a mean of 38.5 and standard deviation of 12.2 and the number of days with a mean of 16.5 and standard deviation of 4.9. These two groups may have different travel patterns of transit uses, as well as different perceptions and willingness-to-pay for transit. Therefore, considering those differences in designing a MaaS plan is necessary to provide the appropriate options that each group needs.

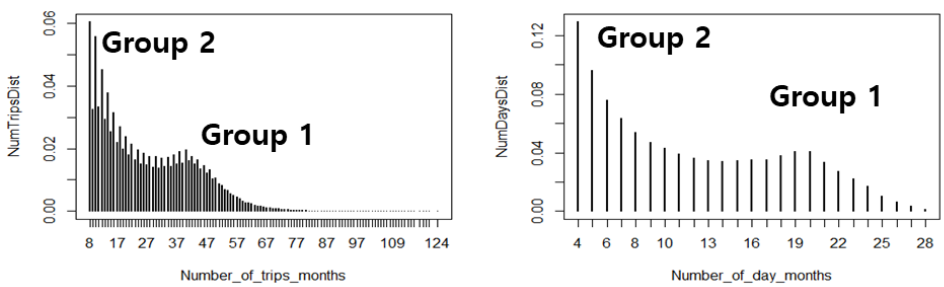


Figure 14. Distribution of the number of trips and the number of travel days within a month

Table 15 shows the trip purpose of transit uses of two groups. While commuting and returning home have a bimodal distribution, other trips show unimodal one. In Group 2, transit is mainly used for other trips rather than commuting, while transit is mainly used for commuting in Group 1.

Table 15 Comparison of the number of trips within a day, the number of travel days within a month collected from one-day, one-week, bi-week, and one month SC data.

Trip purpose	Total		Group 1		Group 2	
	Ratio (%)	Number of trips during a month	Ratio (%)	Number of trips during a month	Ratio (%)	Number of trips during a month
Commute	29.2	7.86	32.0	12.57	21.0	2.92
Other	25.0	6.72	22.4	8.79	32.7	4.55
Returning Home	45.8	12.35	45.6	17.96	46.3	6.45

The effect of individual attributes on classifying these two groups is investigated using a binary logit model. The estimated parameters for individual attributes are provided in Table 16. The estimation results indicate several implications for monthly travel behavior. Younger people tend to be in Group 1, but they would prefer to be in Group B after retirement. Higher-income people are more likely to be in Group 2 than lower-income people. It may suggest that the higher-income people prefer to using a private car, rather than using transit, as expected. People who can't drive tend to be in Group 1, which indicates that Group 1 includes the non-voluntary use of public transportation. Women are more likely to be in Group 1, which is a well-known travel pattern reported in the previous studies (Bhat et al., 2004).

Table 16 Estimation results of the logistic regression model for classifying two groups about monthly travel behavior

Attribute	P(Y=Group2)		
	Category	Estimated parameter	P-value
Age (Reference : > 65)	9~27	-0.16	<0.001
	28~43	-0.27	<0.001
	44~57	-0.15	<0.001
	58~66	0.09	<0.001
Monthly income (Reference : > 10 million won)	> 3 million won	0.17	<0.001
	> 5 million won	0.07	<0.001
Driver license (Reference: Yes)	No	-0.17	<0.001
Car availability (Reference: Yes)	No	0.009	0.15
Gender (Reference: Man)	Woman	-0.34	<0.001

5.4.2. Investigation of customized travel behavior

Individual-specific travel data can be used to understand social phenomena in the public and private sectors. Commercial area analysis is the most important stage for any B2C companies since the demand for the area can be estimated by travel patterns. Transportation is particularly important in this analysis due to its impacts on the location of economic activities as well as their accessibility. The number of activities estimated by the number of trips in SC data would be a proxy measure for indicating demand for the commercial area. I applied the ISC to two commercial analysis, which is a selection of electric scooter location and a selection of hot place (i.e., a place where young people in their 20s and 30s frequently visit at the weekends) location. Office workers or university students of 20~30s at weekends are the main customers for the

electronic scooter market. I estimated the number of trips of the 20s and 30s commuters (i.e., young commuters) and 40s and 50s commuters (i.e., old commuters) to derive the locations where the demand for E-scooters would occur, using ISC data. Table 17 compares the number of trips of young and old commuters. While the top six "dong" districts were the same for these two groups, the rankings of 7 through 9 were different. These three districts would be a promising location for assigning new electronic scooters.

Table 17 Comparison of the number of trips of young and old commuters by districts "dong."

Dong	Gu	20~30s trips	20~30s rankings	40~50s trips	40~50s rankings
Yeoksam 1 dong	Gangnam	77791	1	57550	1
Jongro 1-4 dong	Jongro	57434	2	53860	2
Yeouido dong	Yeungdeungpo	46029	3	52191	3
Myung dong	Joong	38051	4	36814	4
Seocho 3 dong	Seocho	30251	5	34734	6
Sogong dong	Joong	28659	6	36149	5
Seogyo dong	Mapo	28284	7	22425	10
Samsung 1 dong	Gangnam	28020	8	23505	9
Sinchon dong	Seodaemun	27119	9	15583	17
Sajik dong	Jongro	26893	10	27510	7
Gasan dong	Geunchun	24700	11	25073	8
Nonhyun 2 dong	Gangnam	21794	12	22345	11
Apgojung dong	Gangnam	19653	13	15855	15

With the same method, I estimated the number of leisure trips at weekends by young people and old people to derive the hot place. The hot place would be a good location for marketing trendy food, café, and exhibition. Table 18 shows the number of leisure trips of young and old people at the weekend by districts "dong." The places where the difference in the number of trips

between the two groups are large were almost the same for the places investigated by text mining of social network services (Chosunilbo, 2016). This result shows that the simple analysis of ISC can be effectively used for commercial area analysis without a great effort of data collection and mining.

Table 18 Comparison of the number of leisure trips of young and old people at the weekend by districts "dong."

Dong	Gu	20~30s trips	20~30s rankings	40~50s trips	40~50s rankings
Yeoksam 1 dong	Gangnam	9056	3	6360	4
Seogyo dong	Mapo	10410	2	8060	2
Jongro 1-4 dong	Jongro	11612	1	9624	1
Hwayang dong	Gwangjin	4591	10	2619	17
Sinchon dong	Seodaemun	4832	7	3113	12
Myung dong	Joong	8420	4	7049	3
Sajik dong	Jongro	7060	5	6015	5
Hwahwa dong	Jongro	3708	14	2771	15
Samsung 1 dong	Gangnam	3931	12	3208	11

Chapter 6. Conclusion and Future Research

6.1. Conclusion

This study proposed an integrated framework for generating individual-specific activity schedules using DGM. A framework was a data-driven approach that integrates population synthesis and activity scheduling using GM and MM. In addition, the LM combined the NHTS and SC data to complement the quality and utility of the data. Modules making up the framework were evaluated with benchmark models in perspective of the data distribution and the applicability of the ABM.

This study hypothesized that the synthetic NHTS data would represent the population better than the original NHTS data since the generated NHTS data can represent the activities that are missing from the original NHTS data. Evaluation results showed that the synthetic NHTS data were significantly more similar to the population than the original NHTS data. This result showed the DGMs are good at generating data within the manifold that they are trained on, as well as outside the manifold that is included in the population. As the generation, the VAE outperformed the GAN for estimating joint distribution of NHTS. In particular, vanilla VAE outperformed the advanced GAN (i.e., WGAN-GP) due to their intrinsic properties. The CVAE outperformed the full VAE for estimating joint distribution of NHTS, which indicates the effectiveness of factorizing ATA and ISA. Also, this result supports that the LM estimating the ISA conditioning on ATA would be a

reasonable structure. The MM well-performed in creating activity trip-chain using a joint distribution of NHTS in that reasonable trip-chain attributes are generated (i.e., activity sequences, total travel time, and total duration). The LM successfully estimates the ISA of SC data using CVAE. Differences in the distribution of labeled ISA are much lower than those of ATA of SC data and NHTS data. ABM's performance measure shows the promise of labeled ISC data. The labeled ISA shows decent performance for estimating ISA of 1.3km×1.3km spatial resolution. Also, ISA from ISC, NHTS, and Census are evaluated by the district to verify the quality of data. The result shows that the LM can estimate realistic individual attributes, which is enough to continuously monitor the changes in microscopic travel behavior.

The application of this study focusing on the ISC data that are original data obtained from the LM is provided in two perspectives: monthly travel behavior and customized travel behavior. Analysis of monthly travel behavior reveals that travel behavior can be differently observed depending on the duration of the data analysis. Therefore, analyzing only single-day travel data can cause overestimation of transit uses. Monthly travel behavior reveals that the number of trips and the number of travel days has bimodal distributions. These two groups are characterized by ISA of ISC using a logistic regression model. Customized travel behavior is also investigated using estimated ISC. Customized commercial analysis for selecting the location of E-scooter and Hot-place shows that the ISC can represent interesting travel patterns in urban mobility by replacing costly surveys for specific travel patterns.

The impact of this study on other transportation research field would be considerable. The proposed framework addresses the data incomplete issues derived from emerging data. Other emerging data collected from connected vehicles, car navigation, and automatic vehicle identification can be integrated to create more valuable data in managing traffic congestion. Also, the GM produces synthetic data that preserve the relationship between variables but keep data privacy since the agents do not directly correspond to real people anymore. Therefore, it can be used to provide any publicly available datasets to satisfy both data privacy and societal goals.

6.2. Future research

Although the present study proposed an extensive framework for synthesizing individual-specific activity schedules, several points should be considered in future research. First, the application of the framework should be conducted to other emerging datasets to verify the generalized performance. In particular, locational data generated by a mobile device is the most important application subject of future research. These locational data can be used to complement the information of last-mile mobility that are not be collected in smart card, or other mobility services. To elucidate the unobserved behavioral attributes of the locational data, a smartphone-based travel survey would be the most necessary dataset applicable in the near future. Furthermore, mobility trip-chain that are collected from MaaS could integrate all the mobility data, and it will cost-efficiently replace the NHTS. The proposed framework is still

useful for those data to improve the data privacy and applicability of those data.

Second, the integration of modules in the framework would be an important subject to improve the quality of the synthetic data. Specifically, the GM and LM cooperatively learn the joint probability distribution of the SC and NHTS data, which are in the different domains. In computer vision research, image translation has been successfully conducted using a modified GAN structure (Zhu et al., 2017) with unpaired input-output samples (i.e., learning between different domains is possible even if data in one domain is not paired with data in other domains). Those model structures also can be applied to integrate the SC and NHTS data. Also, the GM and MM can be integrated by synthesizing activity trip-chain rather than synthesizing consecutive trips. Although synthesizing activity trip-chain of NHTS data was not considered because it only contains a one-day trip chain, synthesizing long-term travel data that are continuously collected every day could be considered using RNN-based DGM (Lin et al., 2017) and attention network-based DGM (Zhang et al., 2019). To improve the efficiency of the model in the integration stages, reducing high-dimensional data is also important from the modeling perspective. Specifically, dimensionality reduction of location data using an advanced geo-coding method such as positional encoding (Takase and Okazaki, 2019) can be applied to the proposed framework.

Third, the ISC data generated by the LM can represent long-term travel behavior with valuable individual attributes. Monthly or biweekly

travel behavior is key information for designing the MaaS subscription plan, but that information only has been collected using a stated preference survey. Extracting individual-specific long-term travel behavior based on revealed preferences such as SC data would provide new insight for subscription planning and market segmentation.

Reference

Albert, A., Strano, E., Kaur, J. and González, M., 2018, July. Modeling urbanization patterns with generative adversarial networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2095-2098). IEEE.

Alsger, A., Tavassoli, A., Mesbah, M., Ferreira, L. and Hickman, M., 2018. Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, *87*, pp.123-137.

Arentze, T.A. and Timmermans, H.J., 2004. A learning-based transportation oriented simulation system. *Transportation Research Part B: Methodological*, *38*(7), pp.613-633.

Auld, J. and Mohammadian, A.K., 2012. Activity planning processes in the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model. *Transportation Research Part A: Policy and Practice*, *46*(8), pp.1386-1403.

Bantis, T. and Haworth, J., 2017. Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics. *Transportation Research Part C: Emerging Technologies*, *80*, pp.286-309.

Bar-Gera, H., Konduri, K., Sana, B., Ye, X. and Pendyala, R.M., 2009, January. Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board, Washington, DC*.

Ben-Akivai, M., Bowman, J.L. and Gopinath, D., 1996. Travel demand model system for the information era. *Transportation*, *23*(3), pp.241-266.

Bhat, C.R., Guo, J.Y., Srinivasan, S. and Sivakumar, A., 2004. Comprehensive econometric microsimulator for daily activity-travel

patterns. *Transportation Research Record*, *1894*(1), pp.57-66.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Borysov, S.S. and Rich, J., 2019. Introducing super pseudo panels: Application to transport preference dynamics. *arXiv preprint arXiv:1903.00516*.

Borysov, S.S., Rich, J. and Pereira, F.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.

Bradley, M., Bowman, J.L. and Griesenbeck, B., 2010. SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, *3*(1), pp.5-31.

Castiglione, J., Bradley, M. and Gliebe, J., 2015. *Activity-based travel demand models: a primer* (No. SHRP 2 Report S2-C46-RR-1).

Čertický, M., Drchal, J., Cuchý, M. and Jakob, M., 2015, June. Fully agent-based simulation model of multimodal mobility in European cities. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* (pp. 229-236). IEEE.

Choupani, A.A. and Mamdoohi, A.R., 2016. Population synthesis using iterative proportional fitting (IPF): A review and future research. *Transportation Research Procedia*, *17*, pp.223-233.

Deming, W.E. and Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), pp.427-444.

Doherty, S.T., 2000, January. An activity scheduling process approach to understanding travel behavior. In *79th Annual Meeting of the Transportation Research Board, Washington, DC*.

Drchal, J., Čertický, M. and Jakob, M., 2016. VALFRAM: validation framework for activity-based models. *Journal of Artificial Societies and*

Social Simulation, *19*(3).

Drchal, J., Čertický, M. and Jakob, M., 2019. Data-driven activity scheduler for agent-based mobility models. *Transportation Research Part C: Emerging Technologies*, *98*, pp.370-390.

Drechsler, J. and Reiter, J.P., 2011. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, *55*(12), pp.3232-3243.

Drechsler, J. and Reiter, J.P., 2011. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, *55*(12), pp.3232-3243.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L. and Stolfo, S., 2002. A geometric framework for unsupervised anomaly detection. In Applications of data mining in computer security (pp. 77-101). Springer, Boston, MA.

Farooq, B., Bierlaire, M., Hurtubia, R. and Flötteröd, G., 2013. Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, pp.243-263.

Goodfellow, I., 2016. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A.C., 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems* (pp. 5767-5777).

Guo, J.Y. and Bhat, C.R., 2007. Population synthesis for microsimulating travel behavior. *Transportation Research Record*, *2014*(1), pp.92-101.

Hafezi, M.H., Liu, L. and Millward, H., 2018. Learning daily activity sequences of population groups using random forest theory. *Transportation research record*, *2672*(47), pp.194-207.

Hafezi, M.H., Liu, L. and Millward, H., 2018. Learning daily activity

sequences of population groups using random forest theory. *Transportation research record*, *2672*(47), pp.194-207.64

Han, G. and Sohn, K., 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transportation Research Part B: Methodological*, *83*, pp.121-135.

Hermes, K. and Poulsen, M., 2012. A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, *36*(4), pp.281-290.

Jang, E., Gu, S. and Poole, B., 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jung, J. and Sohn, K., 2017. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems*, *11*(6), pp.334-339.

Kim, E.J., Kim, Y. and Kim, D.K., 2020, August. Interpretable Machine Learning Models for Estimating Trip Purpose in Smart Card Data. In *Proceedings of the Institution of Civil Engineers-Municipal Engineer* (pp. 1-22). Thomas Telford Ltd.

Kim, E.J., Kim, Y. and Kim, D.K., 2020, August. Interpretable Machine Learning Models for Estimating Trip Purpose in Smart Card Data. In *Proceedings of the Institution of Civil Engineers-Municipal Engineer* (pp. 1-22). Thomas Telford Ltd.

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D.P. and Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

KTDB (Korea Transport Database) (2016) Household Travel Survey Data. KTDB, Sejong, Republic of Korea. See https://www.ktdb.go.kr/eng/contents.do?key=263 (accessed 17/06/2020).

Kusakabe, T. and Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, pp.179-191.

Liang, Y., Cui, Z., Tian, Y., Chen, H. and Wang, Y., 2018. A deep generative adversarial architecture for network-wide spatial-temporal traffic-state estimation. *Transportation Research Record*, *2672*(45), pp.87-105.

Lv, Y., Chen, Y., Li, L. and Wang, F.Y., 2018. Generative adversarial networks for parallel transportation systems. *IEEE Intelligent Transportation Systems Magazine*, *10*(3), pp.4-10.

Miller, E.J. and Roorda, M.J., 2003. Prototype model of household activity-travel scheduling. *Transportation Research Record*, *1831*(1), pp.114-121.

Miller, E.J., 2017. Modeling the demand for new transportation services and technologies. *Transportation research record*, *2658*(1), pp.1-7.

Mirza, M. and Osindero, S., 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Nowok, B., Raab, G.M. and Dibben, C., 2016. synthpop: Bespoke creation of synthetic data in R. *J Stat Softw*, *74*(11), pp.1-26.

Pritchard, D.R. and Miller, E.J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), pp.685-704.

Rasouli, S. and Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, *18*(1), pp.31-60.

Rasouli, S. and Timmermans, H., 2014. Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, *18*(1), pp.31-60.

Rasouli, S. and Timmermans, H.J., 2014. Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transport and Infrastructure Research*, *14*(4).

Roorda, M.J., Miller, E.J. and Habib, K.M., 2008. Validation of TASHA: A 24-h activity scheduling microsimulation model. *Transportation Research Part A: Policy and Practice*, *42*(2), pp.360-375.

Saadi, I., Mustafa, A., Teller, J., Farooq, B. and Cools, M., 2016. Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, *90*, pp.1-21.

Shen, L. and Stopher, P.R., 2013. A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C: Emerging Technologies*, *36*, pp.261-267.

Smith, L., Beckman, R. and Baggerly, K., 1995. *TRANSIMS: Transportation analysis and simulation system* (No. LA-UR-95-1641). Los Alamos National Lab., NM (United States).

Sochor, J., Arby, H., Karlsson, I.M. and Sarasini, S., 2018. A topological approach to Mobility as a Service: A proposed tool for understanding requirements and effects, and for aiding the integration of societal goals. *Research in Transportation Business & Management*, *27*, pp.3-14.

Sohn, K., Lee, H., and Yan, X., 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems,* pp. 3483-3491.

Sun, L. and Erath, A., 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, pp.49-62.

Sun, L., Erath, A. and Cai, M., 2018. A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, *114*, pp.199-212.

Takase, S. and Okazaki, N., 2019. Positional encoding to control output sequence length. *arXiv preprint arXiv:1904.07418*.

Tanton, R., 2014. A review of spatial microsimulation methods. *International Journal of Microsimulation*, *7*(1), pp.4-25.

van Cranenburgh, S. and Alwosheel, A., 2019. An artificial neural network

based approach to investigate travellers' decision rules. *Transportation Research Part C: Emerging Technologies*, *98*, pp.152-166.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O. and Graves, A., 2016. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems* (pp. 4790-4798).

Wang, Y., Zhang, D., Liu, Y., Dai, B. and Lee, L.H., 2019. Enhancing transportation systems via deep learning: A survey. *Transportation research part C: emerging technologies*, *99*, pp.144-163.

Wen, T.H., Gasic, M., Mrksic, N., Su, P.H., Vandyke, D. and Young, S., 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Xiao, G., Juan, Z. and Zhang, C., 2015. Travel mode detection based on GPS track data and Bayesian networks. *Computers, Environment and Urban Systems*, *54*, pp.14-22.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K., 2019. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems* (pp. 7335-7345).

Yang, L.C., Chou, S.Y. and Yang, Y.H., 2017. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*.

Yazdizadeh, A., Patterson, Z. and Farooq, B., 2019. An automated approach from GPS traces to complete trip information. *International Journal of Transportation Science and Technology*, *8*(1), pp.82-100.

Zhang, H., Goodfellow, I., Metaxas, D. and Odena, A., 2019, May. Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354-7363). PMLR.

초 록

대규모 고차원 통행 자료를 제공하는 새로운 기술들은 교통수요 분석을 위한 데이터 기반 접근방식의 활용을 증대시켜 왔다. 스마트 카드나 모바일 기기의 위치 자료는 개인의 통행 사슬을 연속적으로 기록한다. 하지만 이러한 자료는 통행목적이나 나이, 성별, 소득과 같은 개인 속성이 누락된 불완전 자료이다. 따라서, 신규 기술을 통해 수집되는 자료의 누락된 정보를 추정하는 것은 교통수요 분석, 통행행태 분석 등의 추가분석으로 자료의 활용을 확장하기 위해 필요하다.

현대의 활동기반모형은 현실적인 재현 인구와 그들의 하루 내 활동계획을 핵심 입력 자료로 활용하는 미시시뮬레이션 접근법을 채택하고 있다. 그러므로, 개인 속성과 활동계획을 생성하는 데이터 기반의 접근법은 새로운 기술들로부터 수집되는 귀중한 자료를 활용하기 위해 반드시 필요하다. 전통적으로, 인구 재현과 활동 계획은 서로 다른 접근법을 통해 독립적으로 수행된다. 인구 재현은 인구의 개인속성들의 결합확률분포를 추정함으로써 수행되는 반면, 활동 계획은 전문가가 설계한 여러 개의 연속된 기계학습 기반 이산 선택모형들의 집합으로 수행된다. 따라서, 기존 활동 계획 모형을 새롭게 수집되는 자료에 맞춰 개선하기 위해선 많은 전문가의 노력이 필요하다. 본 연구는 개인 속성이 명시된 활동 계획을 통행 사슬을 기록하는 통행 자료를 활용하여 재현하는 딥생성모형 프레임워크를 개발한다. 제안된 데이터 기반의 프레임워크는 개인 속성과 활동 계획의 결합확률분포를 동시에 추정하여

개인 속성이 명시된 활동계획을 생성한다. 또한, 이 프레임워크는 불완전한 표본 자료의 누락 정보를 추정하여, 완전한 표본 자료와 함께 활용이 가능하다.

제안된 프레임워크를 검증하기 위해, 완전 표본 자료인 가구통행실태조사(NHTS) 자료와 대중교통 스마트카드 (SC) 자료에 적용하였다. NHTS 자료의 추정된 결합확률분포는 현실적인 재현 NHTS 자료를 생성하고, SC 자료의 누락된 개인속성정보를 추정하는데 활용되었다. 이러한 재현 NHTS 자료와 개인 속성이 추가된 SC 자료는 확률분포, ABM으로 활용가능성, 행태분석으로 활용가능성의 3가지 관점에서 벤치마크 모형들과 함께 평가되었다. 평가 결과는 재현된 NHTS 자료가 개인의 활동계획에 대한 결합 확률 분포를 잘 추정함을 보였고, 그러한 성능은 ABM과 행태 분석에 활용되기에 충분하였다. 결과는 또한 데이터 퓨전 기법으로 추정된 개인 속성이 추가된 SC자료가 센서스나 NHTS 자료와 비교했을 때 매우 현실적인 개인 속성을 나타냄으로써, 비용이 많이 소모되는 기존의 통행설문조사를 대체할 수 있음을 시사하였다. 이러한 발견들은 본 연구가 자료의 유용성, 퀄리티, 안전성, 비용 효율성 등을 높여 통행 수요 및 행태 분석 분야에 상당히 기여할 수 있음을 시사한다.

주요어: 활동기반모형, 인구 재현, 활동 계획, 딥생성모형, 데이터 퓨전

학번: 2016-21248