2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Learning Activity Behavior Choice Models Without Personal Data to Generate Behavioral Data for Social Simulations

Asako Yumoto[1], Shinsa Yamaguchi[1], Bin Chen[1], Nozomi Fukuda[2], Tadahiko Murata[2], and Eigo Segawa[1]

Abstract—**This research proposes a method to generate arbitrary activity data for social simulation by utilizing a behavior choice model, learned from synthetic population data and activity data derived from public statistics without relying on hard-to-obtain personal data (actual behavior data). The effectiveness of the proposed method is validated using the activity simulator ActivitySim.**

I. INTRODUCTION

The use of agent-based simulation (ABS) for policy examination of complex social problems has been observed. Although ABS is an effective technique for simulating individual behaviors in detail, using actual personal data (raw data) for simulations—such as data on inhabitants of the target area—is challenging owing to high survey acquisition costs and concerns about personal information protection.

Therefore, a method [1] has been developed to create a virtual population (synthetic population data) that is similar to the real population using public statistical information, particularly national census data [2]. This method generates household and household member (person) data with static human attributes (such as household type, age, sex, and employment status) that remain unchanged daily, while also addressing concerns about personal information protection.

Social simulation of movement requires daily activity data for each individual. Activity data [3] represent the daily activities and movements of an individual as illustrated in Fig. 1. It can be described by a schedule that includes time, location, purpose, and mode of movement. Specifically, it generally comprises a destination trip group (such as commuting, shopping, eating out, and returning home, as shown in Fig. 1) for each movement unit, and a tour (or continuous chain) of these trip groups with the same departure and return locations.

In addition to the static attributes of each individual, such as household composition and employment status, activities are substantially influenced by environmental factors, such as changes in transportation route costs in terms of time and fare. For this reason, activity based simulators has been developed [4] [5] to generate individual activity data by simulating action selection according to individual attributes within a given environmental context. However, these simulators have a limitation; they are required to be pre-trained on a set of action choice models in the simulator, and it is crucial that the training data be actual personal data (including human attribute and activity data).

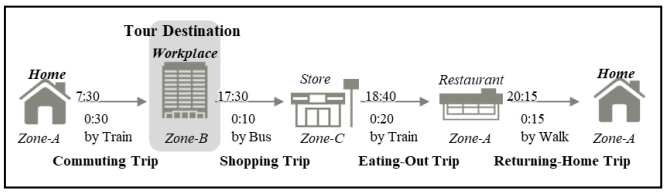

Figure 1. Example of activity data (*commuter tour*)

In this study, we propose a new approach for developing the behavior choice model of the ActivitySimulator by generating training data from synthetic population data and public statistics, without relying on actual personal data. Since the target model uses household information to simulate realistic behavior choices, the generated training data must accurately reproduce both the attributes of individual household members and the overall household composition. Therefore, the proposed approach uses comprehensive census data as a support for activity statistics from sample surveys, which often lack detailed household attributes. The method synthesizes individual attributes and activity data to establish statistically accurate relations, incorporating household composition, and it was developed as part of this study.

In practice, the model group of the ActivitySimulator was trained using the training data synthesized by this method. The activity data generation for test data was then evaluated using the trained model. The results indicated that the model training was effectively accomplished with the synthesized training data. As a result, we confirmed that our approach effectively generates the initial behavior choice model for ActivitySimulation using only synthetic data derived from statistics, without the need for actual personal data.

In this article, we first review related research in Section II. Section III provides an overview of our approach. We use five models related to work tours from ActivitySim [4] as examples of ActivitySimulator. We discuss the human attributes used in the utility of these ActivitySim models, particularly focusing on the behavior choice model, and clarify which human attributes should be prioritized in the training data. Subsequently, we provide a brief overview of the synthetic population data and activity-related public statistics used in generating the training data. We then explain our method for generating the training data and conclude with an explanation of the model learning process using this data.

Section IV presents the movement target areas, including residences, trip destinations, and workplace areas of incoming and outgoing workers in Saiwai Ward, Kawasaki City, as well as the surrounding area of Kawasaki City. Additionally, we describe an evaluation method to assess the effectiveness of the training data generated by the proposed approach. In the

[1] Social digital twin core project, Converging technologies laboratory, Fujitsu Research, Fujitsu Limited, Kawasaki, Japan. asako@fujitsu.com, shinsa.yamaguchi@fujitsu.com, chen.bin@fujitsu.com, eigo.segawa@fujitsu.com
[2] Osaka University, Osaka, Japan. osearchu2@gmail.com, tadahiko.murata.cmc@osaka-u.ac.jp

effectiveness evaluation, we compare the ActivitySim output for test data before and after training with the true values used for training. This comparison assesses whether the output closely aligns with the true values after training. Section V presents the evaluation results, which consequently demonstrate the effectiveness of the training-data creation method using only public statistics by accurately constructing the behavior choice model of ActivitySim for Saiwai Ward, Kawasaki City, Japan. Section VI discusses future developments.

II. Related Framework

A. Method of the ActivitySimulation Using Activity Choice Models

First, according to the concept that the activity of an individual comprises a series of choices regarding activity attributes (such as time, place, purpose, and transportation mode), including decisions about whether to undertake trips and tours as illustrated in Fig. 1, numerous activity-based approaches have been developed to reproduce activity data. These approaches involve creating an activity behavior choice model for each choice according to a behavior choice model. ActivitySimulators, such as ActivitySim [4] and MATSim [5], have been developed and published for this purpose.

Activity data generated using these ActivitySimulators incorporate a behavior choice model that simulates individual decision-making. This behavior choice model allows the simulators to adapt to various situational changes, such as changes in traffic routes and fare adjustments, and to generate more realistic activity data, including variations in activity decisions according to household attributes (for example, having a child in the household). The capability of the behavior choice model also makes the simulators highly compatible with social simulations that aim to project future behavioral changes following policy introductions.

However, many of these simulators are developed by administrative agencies and academic institutions, and it is common to use raw activity data (partially anonymized) from resident surveys obtained directly from administrative agencies as training data for the behavior choice model of the ActivitySimulator, particularly for determining the model parameters [6] [7] [8].

Additionally, there is an approach [9] that generates individual movement demand using an ActivitySimulator, based on anonymous mobile spatial data that tracks each individual's behavior for a day. However, owing to the lack of data, it is not feasible to select activities according to household composition. As a result, this approach is often used solely as a traffic simulator or cannot be effectively applied without separately prepared actual personal data (such as travel history data with household information for calibration).

B. Method Using the Public Aggregate Statistics

In Japan, public statistics on activity data are limited to aggregated statistics including area, time, and daily activity histories (diary data), commonly found in travel and lifestyle surveys, are not revealed. Consequently, a study [10] estimates the distribution of start or end times for various activities based on national statistical information. This study uses aggregate statistics on the proportions of individuals by sex, age group, and job type in each time zone for activities such as commuting, schooling, and sleeping, as reported in the NHK Broadcast Culture Research Institute's National Time Survey [11] (now NHK-TUS). This estimation is useful for synthesizing activity data. However, the activity attributes are not assigned to each unit according to the estimated distribution, and the reproduction of activities for individual and household units is not conducted.

III. Approach and Training-Data Creation Method

A. Approach Overview

For the previous studies discussed in Section II, we propose an approach where behavior choice models are trained using synthesized training data for ActivitySimulators, relying solely on widely available statistical information instead of hard-to-obtain raw activity data. This approach, illustrated in Fig. 2, comprises two main phases. Phase [i] involves model learning and includes two detailed subphases: [i-a] preparation (training-data generation) and [i-b] learning. Phase [ii] focuses on generating new activity data (such as tours and trips) for target environments and inhabitants using the model developed in phase [i].

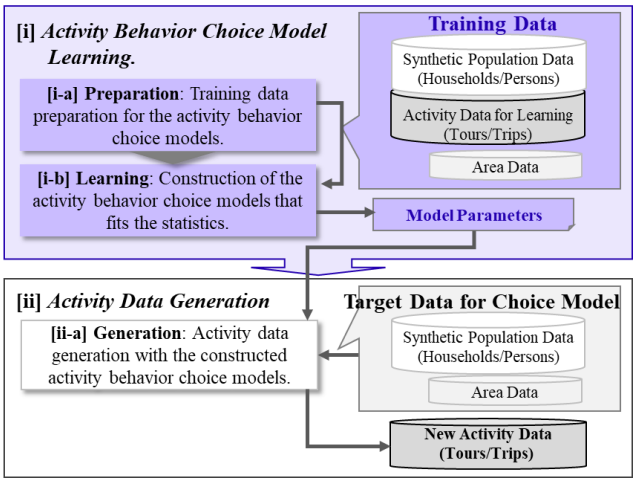

Figure 2. Our approach overview

In this study, the prototype-mtc of ActivitySim, an open-source software from the U.S., was used as the ActivitySimulator and behavior choice model, considering its potential for commercial use.

To assess feasibility, the study focuses on commuting activities (work tours) and assumes that each individual undertakes at most one direct return work tour. The training data are created for a miniature population, smaller than the actual population, to reduce the computational cost associated with learning in phase [i-b].

Table I shows the items for the training data, target data, and new output activity data used in this approach. When using ActivitySim in practice, additional data such as land use information for the target area (including urbanization level and area, as shown in Fig. 2) and transfer cost details between areas (including distance, fare, time required, and number of transfers) are also needed. These data are created using

conventional methods, and the details of their creation are omitted from this paper.

In the learning phase [i-b], five ActivitySim models related to the target commuting tours (Section B) are adapted from the original models trained for the U. S. An initial model of activity action selection for Japan is constructed, and simulating how residents (household members) with specific human attributes perform various activities in different land use environments becomes possible.

B. Used Behavioral Choice Model

ActivitySim estimates each of the activity attributes (such as time, place, purpose, and transportation) involved in trips and tours, as shown in Fig. 1, using 26 discrete choice models. Although these models are likely tailored to reflect American cultural contexts, the ActivitySim prototype-mtc model is used in this study to assess its basic performance in a different country, Japan. Table II provides a summary of the five main models essential for the estimation of work activities by ActivitySim, including the model type and the number of estimated selections. Table III and Fig. 3 illustrate the number of utility expressions for each explanatory variable and show how each attribute influences the choice decisions of each model. Models are listed in the order of ActivitySim estimation.

TABLE II. SET OF ACTIVITYSIM MODELS TO LEARN

Model Name	Model Overview		
	Subject to Presumption	***Type***	***Number of Choices***
[a] auto_ownership	Estimate the number of vehicles owned by households. There are four choices (0/1/2/3/over 3 cars)	MNL	5
[b] free_parking	Estimate the presence of free parking at the workplace. There are two choices (yes/no)	MNL	2
[c] CDAP	Tour type estimation for each household member roughly has three choices (M/N/H). ・M: Mandatorytour (work or school) ・N: Nonmandatorytour(shopping and eating out) ・H: No tour (at home) CDAP: Coordinated daily activity patterns	MNL	3
[d] work_tour_scheduling	Estimate the tour departure time (start) and arrival time (end) of the work tour. There are 190 choices, the combinations with 19 start and 19 end patterns. ・start and end are 5–23 integer hours	MNL	190
[e] tour_mode	Estimate the main means of transportation for the work tour. There are five choices. But when using no learning model, other choices appear, so for the sake of convenience, add the “other” choice, then total are six choices. ・Walk/Bike/Bus (=WALK_LOC)/Train (=WALK_HVY)/Car(=DRIVEALONEFREE)/Others	NL	5(6)

- The main models include the Coordinated Daily Activity Patterns (CDAP) model [c], which estimates whether a household member performs a work tour; the work-tour-scheduling model [d], which estimates the timing of the work tour; and the tour-mode model [e], which estimates the mode of transportation used for the work tour. Additionally, the auto-ownership model [a], which estimates the number of vehicles owned by a household, and the free-parking model [b], which estimates the availability of free parking at a workplace, were included. These models are important for specifying transportation means. Since the destination estimation model for the work tour uses the workplace attribute value from the synthetic population, it is excluded from the model learning process.

- The discrete choice model uses a utility function to express the satisfaction derived from each choice, based on the assumption that each estimation object (household member) selects the option that maximizes the calculated utility. The five models utilize both the multinomial-logit (MNL) model and the nested-logit (NL) model, which is a hierarchical extension of MNL. In MNL, the utility for option *i* for individual *n* is expressed by a utility function (1) which comprises a definite term V_{in}, a linear sum of an explanatory variable x_{ikn} the utility and a weighting coefficient (model parameter) β_k, and a probability term ε_{in} in representing an error. The selection probability $P_n(i)$ of choice *i* of actual individual *n* in MNL is expressed by (2) using (1). The NL is used when alternatives are grouped into a hierarchical structure and are correlated with each other. NL includes a utility term for the entire hierarchical group, which makes it somewhat more complex than the MNL. Despite this complexity, NL, similar to MNL, expresses the influence of explanatory variables on choice as a linear sum of these variables weighted by model parameters. Therefore, the relative impact of different attributes on utility (and choice) can be estimated from the number of utility expressions associated with each explanatory variable.

$$U_{in} = V_{in} + \varepsilon_{in} = \sum_k \beta_k x_{ikn} + \varepsilon_{in} \quad (1)$$

$$P_n(i) = exp(V_{in}) / \sum_j exp(V_{jn}). \quad (2)$$

- Table III presents the number of utility expressions for each type of explanatory variable in the five models. The high number of expressions indicates that both individual household member attributes and overall household attributes substantially influence the selection process. Fig. 3 displays the number of total utility equations across the five models per seven main human attributes and shows that household attributes are almost as prevalent as the personal attributes of household members. Personal attributes of household members appear most frequently in the following order: employment status, workplace, age, and sex. In creating the training data described in Section E, it was crucial to preserve the relation between human attributes and activity attributes as much as possible. This frequency of appearance (or influence on the choice model) was used to prioritize which human attributes should be given particular attention.

TABLE III. NUMBER OF UTILITY EXPRESSIONS FOR EACH TYPE OF EXPLANATORY VARIABLE *(ATTRIBUTE)*

Model Name	**Total**	**Household Attribute**	**Personal Attribute**	**Other Attribute**
		Composition of household members, annual household income, car ownership	Sex, age, employment status, workplace location (duration of each means per time zone)	Residential characteristics, work location characteristics, number of tours related, tour time related, tour type, availability of each mode of transportation.
[a] auto_ownership	29	23	20	15
[b] free_parking	8	5	7	3
[c] CDAP	189	154	189	10
[d] work_tour_scheduling	65	3	15	54
[e] tour_mode	315	69	233	74

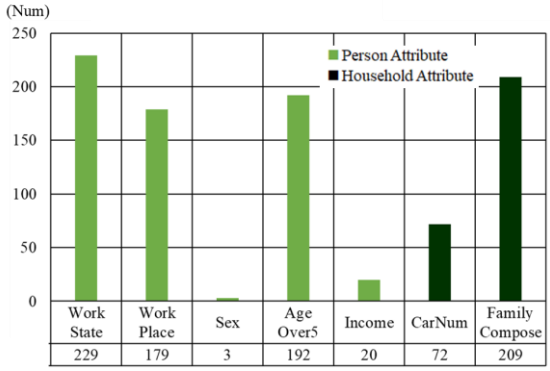

Figure 3. Number of utility expressions for key human attributes in the five models

C. Synthetic Population Data Used to Generate Training Data

The synthetic population data from Osaka University [12], which uses census and work statistics, is utilized for the human attribute data in the training set. This data includes information on every residence place at the town Cho-me code level and encompasses residence place, household type, household ID, household member ID, age, sex, role in the household, and employment attributes (retained only for employed individuals). The employment attributes include engaged industry, employment status, office size, workplace (town Cho-me code), transportation means, and commuting time.

For activity attributes not included in the synthetic population data, the training-data generation synthesizes information according to public statistics, including which household member performs the work tour, the schedule of the work tour, the annual income of individuals and households, and the number of household cars.

D. Public Statistics Used in Training-Data Generation

We use the following four public statistics.

- Tokyo metropolitan area person trip survey [13] (also known as Tokyo PT): This is used as the main source of published statistics for creating learning activity data. These data provide cross-statistics of movement attributes and personal attributes that can be custom-acquired for each fine area, equivalent to the synthetic population of the target area. While this is the Tokyo local statistic, it was selected because it provides finer location resolution and the ability to obtain cross-statistics between more detailed attributes of people and activities.
- NHK TUS[11]: This is the reference statistics for continuous activity execution time (such as working hours at the workplace), which are not available in the Tokyo PT dataset.
- Annual report of employees' Pension Insurance and National Pension Service of the Ministry of Health, Labour and Welfare 2021 [14] and National Tax Agency 2019 tax sample survey results [15]: These are reference statistics used to calculate annual income.

E. Training-Data Generation Method

When creating training data to learn the relation between personal attributes (household member + household attributes) and activities from aggregate statistics, two key requirements must be addressed: 1) "Reproducing individual activity from aggregate activity statistics without diary data" and 2) "Determining which household member corresponds to each individual within a household." Specifically, the challenge of accurately reproducing the relation between household information and activities is a new issue that arises from this approach.

We propose a method for creating training data by synthesizing each individual's activity from activity statistics (Tokyo PT) while utilizing household information from synthetic population data derived from census data. The main innovations are as follows:

- The relation between personal attributes and activities is maintained by associating each personal attribute (place of residence × household type × sex × age group × employment status) with each household member in the synthetic population data and with the synthesized activity content from Tokyo PT. To ensure consistency, the correspondence is made at the household level rather than the individual level, preserving household information such as household member composition.
- The distribution of human attributes notably varies between the census, which surveys the entire population and serves as the basis for the synthetic population, and the Tokyo PT, which is an extended sample survey. For example, the number of full-time workers in Tokyo PT is 1.5 times higher than in the census (in Saiwai Ward, Kawasaki City). Consequently, even when working with a miniature population for training (for example, 10% of the total population), these differences must be accounted for. When working with a miniature population (for

example, 10% of the original number), the likelihood increases that the attributes of all household members in the synthetic population will not match those in the Tokyo PT group and cannot be assigned. In such cases, the verification of matching personal attributes is progressively reduced and assigned at the household level. To mitigate the impact of these mismatches, the omission of verification is prioritized according to the frequency of utility equations (Fig. 3) in the ActivitySim model being learned. This approach minimizes the effect of discrepancies caused by the omission of attribute verification during model training. This approach is an innovative method for generating training data that minimizes the impact of discrepancies caused by omitted verification in the learning model.

Table IV shows the flow of training-data generation.

If necessary, household member attributes are matched by omitting confirmation in the order of gender, age group, and employment status. The omitted confirmations are then overwritten with attributes from the synthetic population to maintain the integrity of the household composition (Table IV(b)(c)). In this process, children under five years old, who are not included in the PT statistics, are sourced from the combined synthetic population.

Afterward, the number of commuting trips for each household is determined according to whether they commute, using the number of trips for each activity type and personal attribute from Tokyo PT (Table IV(d)). Additionally, verification is carried out to check whether the workplace of the synthetic population is outside the simulation analysis target area (discussed in Section VI.A). If the workplace is outside the target area, it is tentatively assigned randomly within the target area. In that case, the mode of transportation is determined using a heuristic weight rule according to the distance between residence and workplace, and the travel time (commuting time) is set according to the average travel speed for each mode of transportation (Table IV(e)).

Furthermore, the details of the commuting trip, including the schedule, are determined, and the return trip is added to create the work tour (Table IV(f)–(h)). For the work schedule, the arrival time at the workplace is established using the distribution of starting times (by hour) according to residence and working status from Tokyo PT. The working hours at the workplace are determined using the average working time from NHK-TUS, supplemented by a random value representing standard deviation (in hours). The schedule of the work tour is then finalized by incorporating the commuting time from the synthetic population data.

Finally, household attributes such as annual income and the number of cars owned are calculated according to the annual income of household members and the number of members commuting by car (as indicated by the commuting means in the synthetic population data) (Table IV(i)). The annual income for household members is determined using the average annual income and average pension amount, stratified by employment status and sex, along with the standard deviation values from published statistics. These values are then applied using a standard deviation random method.

TABLE IV. FLOW OF TRAINING-DATA GENERATION

	Outline of Procedure	***Relevant ASIM Models***	***Statistics Used***
(a)	Preparation of household members (individual household member Tokyo PTs) in line with the Tokyo area PTs.		• Tokyo PT custom statistics: resident population [persons] = (place of residence (total base zone)) × (sex-2) × (age group-17) × (employment status-5) × (household composition-5)
(b)	Household/household member assignment of synthetic population data to individual household member Tokyo PTs.		
(c)	Attribute merging of individual household member Tokyo PT votes (overridden by synthetic population attributes).		
(d)	Commuting trip number (commuting or not) assignment of individual household member Tokyo PTs (Trip chain creation).	[c]CDAP	• Tokyo PT custom statistics: intensity [Trip] = (Activity type) × (place of residence (total base zone)) × (sex-2) × (age group-8) × (employment status-5) × (household composition-5)
(e)	Check Trip destination (place of work)/transportation/travel time for individual household member Tokyo PT.	[b]free_parking, [e]tour_mode_choice	
(f)	Assignment of trip time schedules (start/end of work/time of attendance) for individual household member Tokyo PTs.	[d]work_tour_scheduling	• Tokyo PT Other Statistics 2-2: Population statistics by work type by starting time [persons] = (place of residence (Total base zone)) × (time zone-24) × (work attribute-6) • NHK-TUS: Duration of work activity execution [hours] = mean (Ave) and standard deviation (SD)
(g)	Assignment of homebound trips to individual household member Tokyo PTs.		
(h)	Creation of Tour (commuting) from the trip chain of individual household member Tokyo PTs.		
(i)	Assignment of other attributes (household income, number of cars owned) to individual household member Tokyo PT.	[a]auto_ownership	• Annual Report on Employees' Pension Insurance and National Pension Services: average annual income [yen].National Tax Agency 2022 tax return sample survey results: average annual income [yen].

F. Activity Behavior Choice Model Learning

Using the training data created by the proposed method (E.), model learning is conducted for each of the five ActivitySim models. Parameter learning was performed by maximizing the log-likelihood function (3), which assumes a Gumbel distribution, a standard approach for likelihood optimization, to satisfy (4). The scipy library's SLSQP algorithm was used for optimization. Most model parameters for each explanatory variable, originally provided with ActivitySim as U.S. parameters, were adjusted to better fit the Japanese training data.

$$LL(\theta) = \sum_{n=1}^{N} \sum_{i} y_{in} \log P_n(i|\theta) \quad (3)$$

$$\frac{\partial LL}{\partial \theta_k} = 0, \forall_k \quad (4)$$

P_n ($i|\theta$): the probability of observing the data sample i given parameters θ

y_{in}: an indicator variable that equals 1 if individual n is chosen given choice i, and 0 otherwise.

IV. Experiments

As a trial application of the technology developed in Section III E, training data for the work tour of residents in Saiwai Ward, Kawasaki City, was created, and the ActivitySim activity behavior choice models were trained using this data. This section describes the analysis target area for the trial application and the evaluation methods used.

A. Prototype Target Area

ActivitySim prototype-mtc models movement activities (trips and tours) within a predefined analysis area known as Transportation Analysis Zones (TAZs). For this study, 52 TAZs around Kawasaki City were utilized (Fig. 4), focusing specifically on the commuting tour data for Saiwai Ward, Kawasaki City (TAZs 11-13 in Fig. 4).

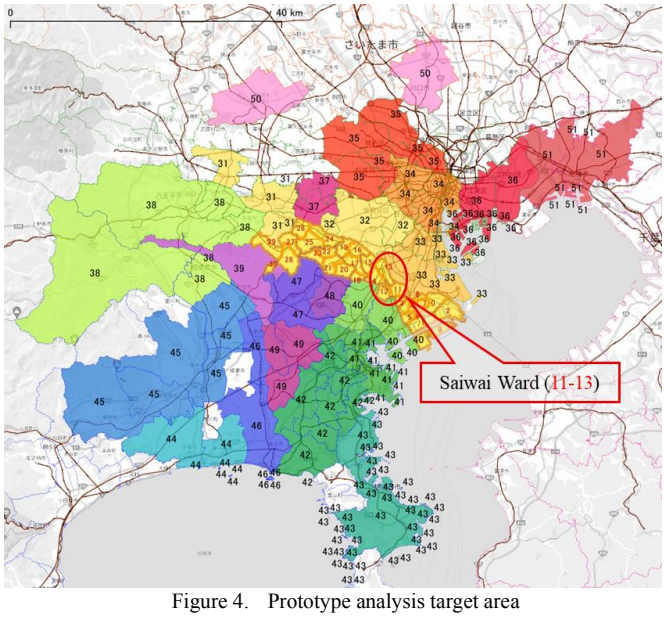

Figure 4. Prototype analysis target area

The TAZs were established by manually splitting and combining 70 municipalities, which represent 80% of the incoming and outgoing workers in Kawasaki City. These municipalities were selected during the generation of workplace attributes for the synthetic population. Kawasaki City itself is divided into finer TAZs (marked in red letters in the figure; TAZs 1-31), while the surrounding areas are grouped into slightly larger TAZs (marked in black letters in the figure; TAZs 32-52) to account for train lines and the challenges associated with commuting by car.

The study data focused on households and household members residing within the TAZs. This included all household members living in Saiwai Ward, Kawasaki City, as well as workers and their cohabiting household members who commute to Saiwai Ward.

B. Evaluation Index

To evaluate the training-data generation method developed for the proposed approach (Fig. 2), we assess whether the ActivitySim model can accurately adapt the American model to a Japanese context using the corresponding training data. The evaluation involves comparing the estimated activity data (output 2 in Fig. 2) generated by the model before and after training with the true values used for learning (training-data activity data) for a test dataset of household and household member data that differs from the training dataset. In the test evaluation dataset, the population size is 100% (in contrast to the 10% used for training data), and the synthetic population data is different because of using another initial value used during random number generation. However, the same area data used during the training phase is applied in the test evaluation.

The evaluation involves assessing model performance compared with the estimated results before and after learning the true values. This is done by analyzing the distribution of results for each of the five models, using histogram distribution evaluation.

The evaluation index values are based on the true values. The false estimation rate (falseR [%]), the average false estimation rate (meanFalseR [%]), and the KL-divergence value (KL($P \parallel Q$)) are used for assessment, where KL-divergence is computed with Laplace correction as described in (5) to (9). Laplace correction adjusts the probability distribution by adding 1 to the count in each bin. For all evaluation values, a value closer to 0 indicates better performance, meaning the output of the model is closer to the true values.

In equation (5), choice k corresponds to each bin of the histogram. When a data point is misestimated, the count in the true value bin is decreased by 1, and the count in the error estimation bin is increased by 1. The number of changes in each bin is then multiplied by 0.5.

$$\text{falseR}[\%] = \sum_{k=1}^{n} 0.5 \times abs(S_k/m - T_k/M) \quad (5)$$

$$\text{meanFalseR}[\%] = \text{falseR}/n \quad (6)$$

$$\text{KL}(P||Q) = \sum_{k=1}^{n} (P(k)\log(P(k)/Q(k))) \quad (7)$$

$$Q(k) = (T_k + 1)/(M + n) \quad (8)$$

$$P(k) = (S_k + 1)/(m + n)\,. \quad (9)$$

M: Total number of training data (true value)

m: Total number of test evaluation data

k: Choice (bin),

n: Total Number of choices,

S_k: Number of test estimates of choice k,

T_k: Number of true value data of choice k,

$Q(k)$: Probability distribution of the true value of choice k,

$P(k)$: Probability distribution of the test estimation of choice k

V. Result and Discussion

Table V presents the evaluation results, while Fig. 5 displays the histograms of the estimation results for each model used in the evaluation.

The upper row of Table V shows the reference values for the untrained U.S. model, while the lower row displays the values for the trained Japanese model. In Fig. 5, the horizontal axis of the histogram represents each bin (or choice), and the vertical axis indicates the percentage of all estimated targets (such as household members, and tours). "GT" refers to the training data.

The error rates for the five models in Table V were relatively high for the [d] and [e] models. For the [d] model, the high error rate is owing to numerous choices, while for the [e] model, it stems from the incomplete cost table of transportation modes used in the simulation, which lacks bus route information spanning both inside and outside Kawasaki City, resulting in fewer bus choices. Despite these issues, the KL-divergence values are close to 0, and the evaluation values show a notable improvement compared with the U.S. model with untrained parameters. This indicates that the training data has led to effective learning.

Thus, it was confirmed that the proposed method, which learns the activity action choice model using training data derived solely from statistics without relying on raw activity data, was effective. This is evidenced by the fact that the estimation results for the five evaluation models were sufficiently similar to the true value data.

TABLE V. Overall Distribution Evaluation

Model Name	Choices *(bin) Number*	*falseR [%]*	*meanFalseR [%]*	*KL-Divergence*
[a] auto_ownership	5	(54.79) 1.22	(10.96) 0.24	(0.86829) 0.00034
[b] free_parking	2	(24.01) 1.23	(12.01) 0.61	(0.30942) 0.00358
[c] CDAP	2	(41.14) 1.42	(13.71) 0.47	(1.20515) 0.00018
[d] work_tour_scheduling	190	(48.75) 9.87	(0.26) 0.05	(0.31032) 0.02760
[e] tour_mode	5	(67.18) 9.34	(11.20) 1.56	(1.65253) 0.01388

The values in () are the reference values for the untrained U.S. model.

VI. Conclusion

Herein, we propose a method for generating training data for ActivitySimulators using synthetic population data derived from public aggregate statistics. This method for preparing training data can also be used to generate activity data for various activities beyond commuting, such as shopping and dining out. With the utilization of Tokyo PT statistics—such as origin-destination data by mode of transportation, the number of trips departing and arriving, and trip generation by previous and next trip purposes—it is possible to create more accurate training activity data, including trip chain orders and activity transition probabilities.

In the future, we aim to enhance the accuracy of activity execution locations and transportation modes using this development method. We will expand the range of supported tour types and activity behavior choice models by incorporating additional activity types. Additionally, we plan to adapt the models of ActivitySim, which currently include U.S.-specific elements such as school transportation, to better fit the context of Japan. We will also calibrate these models using observed data to improve their alignment with real-world conditions. In particular, model calibration focuses on refining initial models, which are trained with data generated from statistical sources, to better reflect real-world conditions. Our goal is to develop a more effective calibration method that can be applied in scenarios and locations where statistical data is not readily available.

References

[1] T. Murata, T. Harada and D. Masui, "Comparing Transition Procedures in Modified Simulated-Annealing-Based Synthetic Reconstruction Method without Samples", SICE Journal of Control, Measurement, and System Integration, Vol.10, 2017, pp.513-519.

[2] https://www.stat.go.jp/data/kokusei/2015/

[3] https://www.tokyo-pt.jp/static/hp/file/pr/pr_tokyoact.pdf

[4] https://ActivitySim.github.io/

[5] https://www.matsim.org/

[6] I. V. de Lima, M. Danaf, A. Akkinepally, C. L. De Azevedo, and M. Ben-Akiva, "Modeling Framework and Implementation of Activity- and Agent-Based Simulation: An Application to the Greater Boston Area", 2018

[7] N. A. Khan, H. Shahrier, "Validation of an activity-based travel demand modeling system", 2021.10

[8] M. Bradley, J. L. Bowman and B. Griesenbeck, "SACSIM: An applied activity-based model system with fine level spatial and temporal resolution", Journal of Choice Modelling, 2010, vol. 3, pp. 5-31,

[9] A. Bassolas, J. J. Ramasco, R.Herranz, and O. G. Cantu-Ros, "Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona", Transportation Research Part A: Policy and Practice, Vol. 121, p. 56-74, 2019.

[10] M. Ichikawa, K. Komori, and J. Xue, "Analysis of Life Behavior Data for Proposing Standard Life Behavior Model in Social Simulation Models", SICE, Socsys 12th, 2017, pp.188-194, (in Japanese)

[11] https://www.nhk.or.jp/bunken/yoron-jikan/

[12] https://www.sde.cmc.osaka-u.ac.jp/SyntheticPopulationE/

[13] https://www.tokyo-pt.jp/person/01

[14] https://www.mhlw.go.jp/topics/bukyoku/nenkin/nenkin/toukei/nenpou/2008/dl/gaiyou_r03.pdf

[15] https://www.nta.go.jp/publication/statistics/kokuzeicho/shinkokuhyoh on2019/pdf/gaiyo.pdf

TABLE I. DATA ITEMS OF OUR APPROACH

Data	*Items in Training Data Created in [i-a]*	*Items in Target Data for Choice Models in [ii]*	*Items in New Creation Data (by Simulation) in [ii]*
Households	Same items as Target Data for Choice Models.	Household ID, residence zone ID, annual income, NumID of household members, household type, number of cars of households, number of workers in households, etc.	Same Items as Target Data for Choice Models. In addition, use the following: Income class, number of drivers/not-work/not-student under 16 age/household members with outgoing, live in rural/urban, percentage of households securing a car for commuting, Travel Num of adult/student/preschool pupil, number of tours with multiple household members.
Persons	Same items as Target Data for Choice Models. In addition, use the following: Whether or not there is free parking at the place of work.	Person ID, household ID belongs, ID in household, sex, age, employment status, student type, person type (mix type of student, employment, retired, etc.), school zone ID, workplace zone ID	Same Items as Target Data for Choice Models. In addition, use the following: School type, distance to workplace, weather the workplace is central business area, area type of workplace, time required to drive to/from work, CDAP, tour number of each activity, number of trip, whether the trip is for work or school.
Tours	Tour ID, household ID, ID of representative household member, tour type, asif mandatory, destination zone ID, origin zone ID, start time, end time, means of transportation (mode)	—	Same Items as Training Data. In addition, use the following: Number of all tours the person did, tour ID in all tours, number of stops, number of sub-tours at work.
Trips	Trip ID, person ID, household ID, tour ID, asif outbound, trip purpose, destination zone ID, origin zone ID, start time, end time, means of transportation (mode)	—	Same Items as Training Data. In addition, use the following: Trip sequence ID by round trip, total number of trips by round trip.

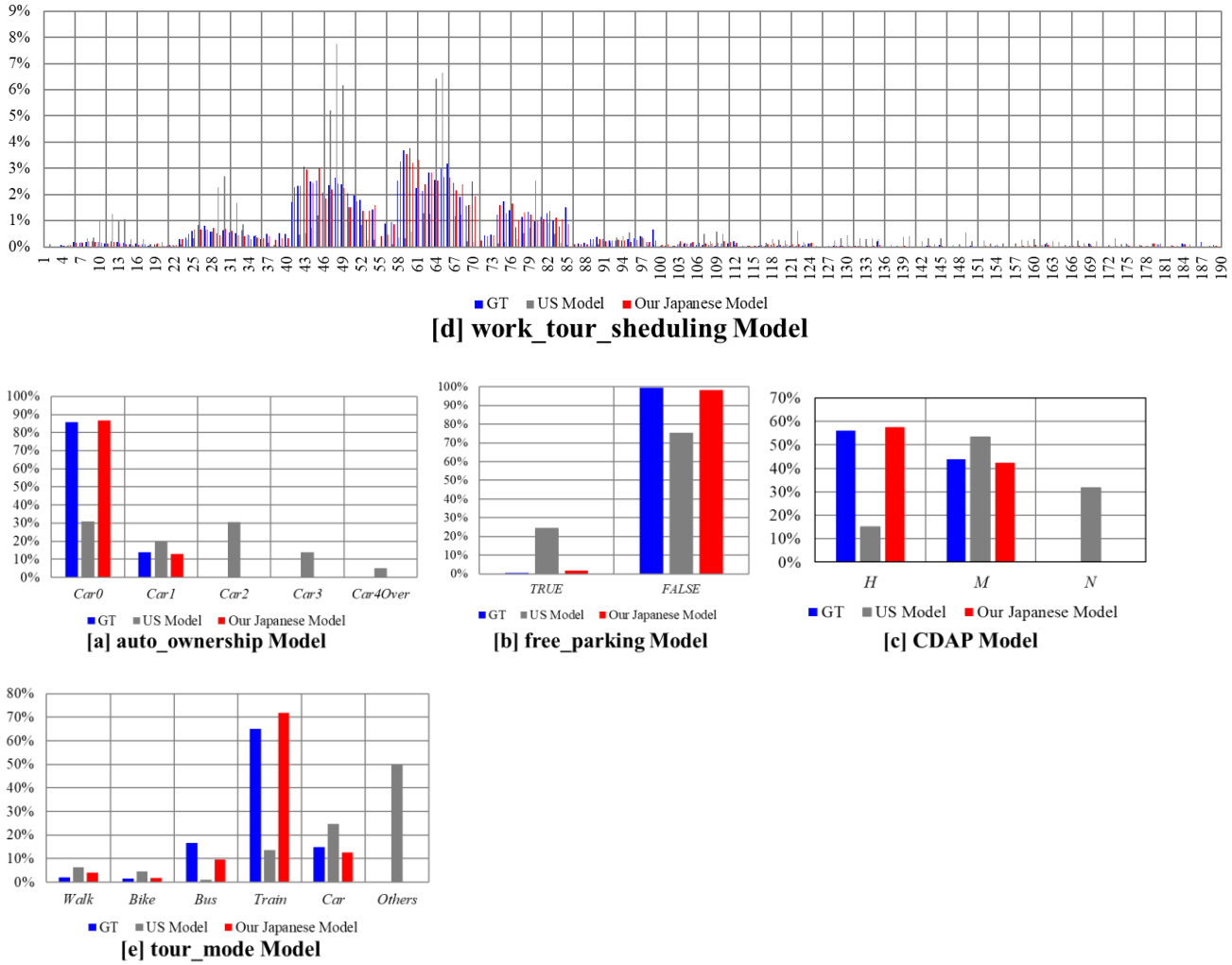

Figure 5. Comparison of Simulation Results of Five Models. (Estimated percentage histogram for each choice. GT is training data.)