3-dimention pedestrian route choice models with cross entropy methods using multi sensor observations

INU021004

Binchang Shen
shen@bin.t.u-tokyo.ac.jp

Takumi Suga
suga@bin.t.u-tokyo.ac.jp

Eiji Hato
hato@bin.t.u-tokyo.ac.jp

Abstract

This paper constructs a three-dimensional (3D) route choice model based on an adjusted RL model and observation model. In order to reproduce the pedestrian trajectory in the 3D space with precision, we designed an observation model using machine learning algorithms and proposed a Data Fusion (DF) based location strategy by integrating classifiers trained from multi-sensor data set to address with signal attenuation issue caused by the environment and access point (AP) distribution. In addition, the Recursive Logit (RL) model is utilized to build a route choice model on a choice-stage network (CSN). Since the output of the observation model is a link set in the form of probability, we introduce cross-entropy instead of the likelihood as the objective function for parameter estimation. A case study at Shibuya Station shows that our model is practical in complex 3D spaces. Pedestrians' route preferences reflected in the model can provide a reference for future research on pedestrian behavior modeling and network optimization.

1. INTRODUCTION

With the rapid development of city areas, growing requirements involving urban design and transportation planning have increased expectations for accurate traffic information and modeling. While the research and application in road traffic networks have made many breakthroughs, the attention and exploration of the urban pedestrian network are still limited (*1*). There are many results of behavioral and network models in the study of automobile traffic and public transportation, and some of the results of Daganzo's logit model (*2*) and Sasaki's Markov model (*3*) may be applicable to the study of pedestrian traffic. However, there are some issues in applying these models. GPS data, which has paved the way for the analysis of automobile traffic, has limitations in its application to pedestrian traffic. It is challenging to observe movement inside buildings in a three-dimensional (3D) urban space. Although information from multiple sensors is available, there are examples of research by Hato (*4*) on how to combine them. Still, it is safe to say that there is a strong need for new models that incorporate machine learning results. And while the decision-making of pedestrians' path-choice behavior is sequential, with a high degree of freedom, it is globally affected by spatio-temporal prism constraints (*5, 6*), so advanced models that take this into account are needed. In particular, the studies on pedestrian route choice behavior under a complicated scenario are not widely available.

At the urban scale, the pedestrian network is denser and with higher resolution than the vehicle network. It sometimes takes on a 3-dimensional form (ex, in transport hubs and some pedestrian facilities). From the micro-view, pedestrians are faced with more frequent path selection and usually with more alternatives. All of the above characteristics lead to higher requirements for pedestrian location accuracy and computation of pedestrian modeling. For micro-scale pedestrian modeling, trajectory monitoring can be well achieved using video (*7*) and various sensors. Nevertheless, when it comes to an entire pedestrian network, such as in a city center district, a position solution that considers precision, scale, and cost remain to be discussed.

In recent years, machine learning (ML) has a good performance in dealing with pattern identification problems on big data and has been widely applied in transportation (*8–10*). This research incorporates machine learning into pedestrians' location identification and constructs a pedestrian model on a 3-dimensional(3D) urban network basing on the location identification result. The overall framework of this study is shown in Fig.1. This study compares several ML approaches and their performance on both WiFi and GPS data set. On this basis, a data-fusion-based location scheme, which takes into account both cost performance and precision, is proposed for better inference. In addition, we replace the log-likelihood function with a cross-entropy function in parameter estimation and confirm the potential of recursive logit (RL) model (*11*) on a probabilistic link set. We apply the proposed framework to the actual sensor records collected in the Shibuya station area. The result shows that it can be an effective tool for the network design of a pedestrian network containing a 3D structure. The overall framework of this study is shown in Figure 1

We organize the paper into six sections: Section 2 details a literature review. In Section 3, we introduce the approaches applied to location identification. Section 4 presents a parameter estimation method using a probabilistic link set. Section 5 describes a case study basing on a survey conducted in the Shibuya station area, and Section 6 analyzes the result with a conclusion.

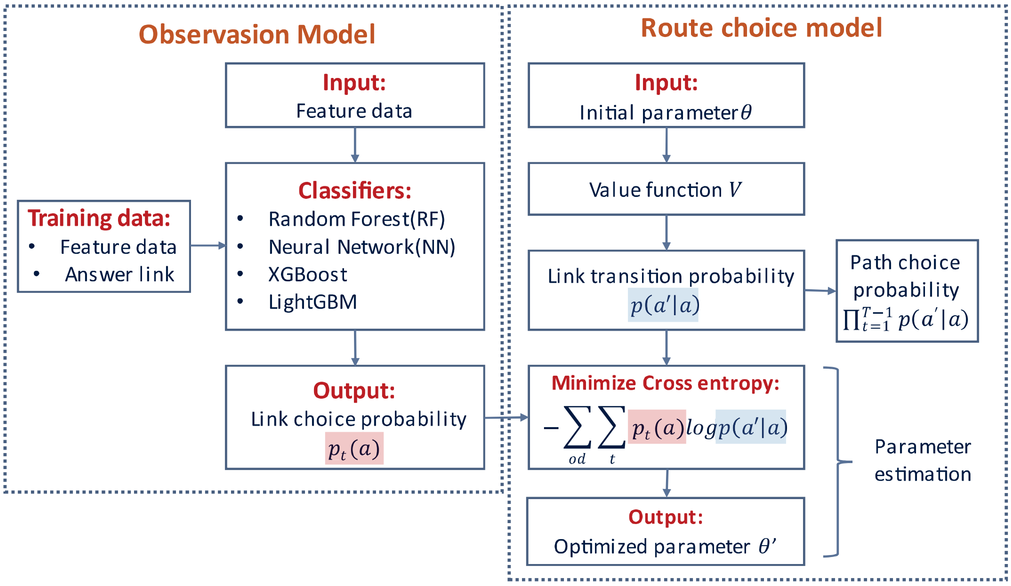

FIGURE 1 Framework of Model Systems

2. REVIEW

2.1 Observation methods

2.1.1 sensor measurement

Given the nature of pedestrian network: high density and diverse branches, accurate and cost-effective positioning technique is essential for establishing effective pedestrian models. Therefore, compared with traditional data collection methodologies such as self-report and stated preference surveys, sensor techniques include Global Positioning System (GPS) traces, Light Detection and Ranging (LiDAR) sensing, Wi-Fi, RFID, Bluetooth sensors, and cameras are more suitable for pedestrian route measurement (*1, 12–17*). On the other hand, given their current cover rate in pedestrian infrastructures, the use of high-precision sensors, such as cameras, usually leads to a tradeoff between accuracy and coverage, not to mention the restrictions due to privacy issues (*13*). Considering the above content and the popularization of smartphones, there is a broad application of GPS and Wi-Fi probes in research and application for the low cost, availability, operability, and globally unified standard (*1, 13, 15, 18*).

The GPS-enabled smartphones are typically accurate to within a 5 meters radius on a wide street. However, compared to its success in open space, GPS signal can be vulnerable to jamming or even wholly blocked under a complicated street environment condition, which leads to measurement error and accuracy deterioration (*1, 12, 13*). On the other hand, the Wi-Fi probe shows an exemplary detection performance on indoor spaces, but the reliability relies on the pre-existing infrastructures and pre-survey on access points (APs) location (*15*).

2.1.2 data fusion and map matching

To improve positioning performance on the entire network, Data Fusion (DF), a collection of techniques that combine multiple sources to achieve better accuracy, can be applied (*8, 19*). Another common position technique is Map matching (MM), which matches the serial geographic coordinates to transportation networks. Many researchers have utilized one or both to contribute to location identification. 1988, Krakiwsky et al. (*20*) developed a Kalman filter to integrate dead reckoning, MM, and GPS, which can estimate the state of a dynamic system from a series of measurement data containing statistical noise and other inaccuracies with the measurement bias caused by signal blocking. Recently, Danalet et al. (*13*) incorporate the Bayesian approach to merge the prior information of the potential activity-episode location to make up for the scarce data. Focusing on the fact that the spatial characteristics within a link tend to be uniform, Oyama and Hato (*1*) introduced a link-based route measurement model and proposed a framework that sequentially determines the links.

Traditionally, the MM process is unidirectional (solid line connection); in other words, it does not have any effect on DF; whereas, MM can be bidirectional (both solid and dashed line connection) for the map can be regarded as an information source and calibrate DF (*19*). In this research, we use a DF-based location scheme, as shown in Figure 2

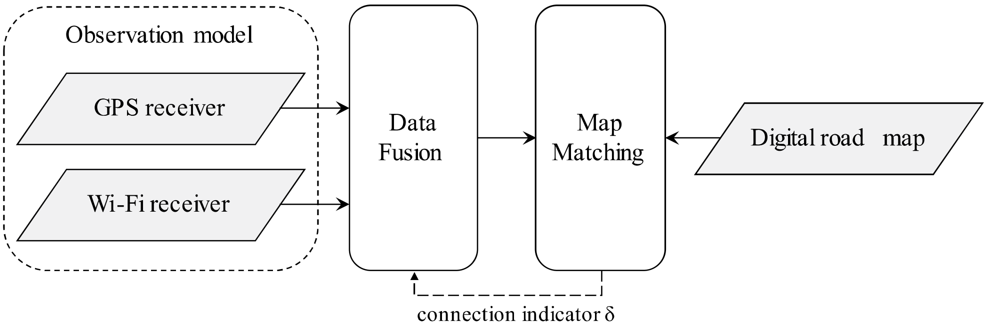

FIGURE 2 a data-fusion-based location scheme

2.2 Route choice modelling

Like other choice problems in transportation fields, such as travel mode selection, the essence of route choice is choosing a specific route with a set of alternative routes (*9*). Most existing route choice models, such as C-Logit (*21*), Path Size Logit (*22*), and Logit mixture (*23*), deal with route choice behavior in deterministic networks, ignoring the dynamics of route choice behavior. Although some previous studies have focused on route choice behaviors in a dynamic network, the alternative routes are path-based (*24*). In other words, these models are basing on the hypothesis that travelers have a global spatial cognition over networks, which limits the applicable traffic scenarios and requires pre-sampling of paths. Sasaki (*3*) proposed an absorbing Markov chain model for traffic assignment, which first incorporates the link transition probability into the traffic assignment model and avoiding explicit path enumeration. Then, The Recursive Logit (RL) model proposed by Fosgerau et al. (*25*) introduces the Markov process into the route choice context, constructs a dynamic discrete choice framework to model route choice as a sequence of link choices. The RL model assumes the traveler always chooses the following link that maximizes the sum of the instantaneous utility and expected downstream utility at each state.

Researchers have proposed extensions to make the RL model framework work better in actual traffic scenarios: The Nested RL model (*26*), cross-nested (RCNL) model (*27*), and mixed recursive logit (MRL) model (*28*) relax the independence from irrelevant alternatives property of logit model, allowing for correlation structure among path alternatives. On the other hand, the DRL model (*11*) introduces a discount factor of expected future utility to generalize travelers' decision-making dynamics. In addition, Oyama and Hato (*5, 6*) incorporated the concept of prism constraints and choice-stage into the route choice model and develop traffic assignment on the choice-stage structured network (CSN). This extension excludes the unworkable routes for more efficient computation and minor requests for memory capacity, making the traffic flow assignment on a high-resolution network possible.

2.3 Estimation approaches

Several approaches have been proposed to estimate dynamic discrete choice models, most of which are two-step estimators, such as the nested fixed-point algorithm (NFXP), Conditional Choice Probability (CCP) estimator, and the nested pseudo-likelihood (NPL) algorithm. A review of estimators of dynamic discrete choice structural models can be found in Aguirregabiria and Mira's work (*29*). For maximum likelihood estimation of single-agent dynamic discrete choice model, Rust (*30*) presented NFXP, which consists of an outer BHHH optimization algorithm and an inner fixed-point algorithm; the implicit value function is solvable using backward induction in the inner loop when the model has a finite horizon. The NFXP model have already been used in the context of route choice modelling, e.g., recursive discrete route choice models (*31*). The main limitation of NFXP is its algorithmic complexity since it requires solving the dynamic programming (DP) problem for each trial value. Hotz and miller (*32*) observed that the repeated solution of DP problem can be avoided and proposed the CCP estimator, which achieves a significant computational gain at the cost of efficiency. Like many other two-step estimators, CCP is not asymptotically efficient and has finite sample bias. As an extension of CCP estimator, the NPL algorithm (*29*) was proved to improve the asymtotic properties and reduce the finite sample bias significantly.

On the other hand, Su and Judd (*33*) proposed a constrained optimization strategy for estimation, referred to as the mathematical programming with equilibrium constraints (MPEC) approach. MPEC can be regarded as an alternative computational algorithm to NFXP for implementing the same statistical estimator at a lower computational complexity.

This paper aims to model pedestrian's route choice behavior in a 3D urban space. We take an FD location strategy that integrates the spatial information to check connection conditions in the pedestrian network and realize precise location identification. Basing on the location result, We adopt RL model on CSN for modeling and utilize the MPEC method to estimate the parameter vector.

3. LOCATION IDENTIFICATION

In this section, we propose a link-based observation model incorporating machine learning for location identification.

3.1 Data structure

In this study, we use two sensor sources for link detection: GPS and Wi-Fi. Let $\hat{m} = (\hat{x}, \hat{\tau})$ denote a vector of measurements. The GPS measurements at time t, $\hat{x}^t_G = (x^t_{lat}, x^t_{lon}, x^t_{alt})$ is a combination of coordinates and altitude; while the Wi-Fi measurements at time t, $\hat{x}_W = (x_1^t, ..., x_N^t)$ is a vector of detected APs at time t (if the device can detect n-th AP at time t, then $x_n^t = 1$, otherwise $x_n^t = 0$), N is the total number of APs in the entire network. $\hat{\tau} = (\tau_1, \dots, \tau_N)$ is a vector of measurement timestamp. The prediction results of both two types of data are in the form of a vector $y^t = ((a_l, p(a_l \mid x_1^t)), \dots, (a_{l'}, p(a_{l'} \mid x_N^t)t))$, containing the class label (link ID) $a_l \in A_l$ and posterior probability of putting the input data x_t. The data set for training the model contains feature array x and link answer y, we define it as $Z = (x_1, y_1), \dots, (x_n, y_n), \dots, (x_N, y_N)$.

3.2 Machine learning methods

Machine learning is a powerful method that automatically analyzes data to obtain patterns and uses statistical patterns to generate a high-quality prediction for unknown data. Machine learning models can be divided roughly into two types: supervised learning and unsupervised learning. The former learn underlying distribution laws of the pre-labeled objects to produce an output, while the latter find the natural grouping from unlabeled data (*8*). This study uses supervised learning algorithms to build a 3D position observation model, which is essentially a multi-classification model. To determine the classifier that can provide the most accurate prediction, we adopted the following four classifiers and compared them in terms of accuracy and efficiency on both GPS and W-Fi data sets.

Random forest

Random forest (RF) was developed by Breiman (*34*) from his bagging idea. RF is an ensemble classification and regression method based on tree-like structures (decision tree) at the training level. The RF classifier applies a “majority voting” mechanism to combine each tree’s classification results.

Neural network

Neural Network (NN) is a mathematical model that imitates the systems of neurons to recognize underlying relationships in a data set. The artificial neurons form the network through the connections of synapses. The strength or amplitude of a connection (also called synaptic weight) changes while training, lead to a different level of trigger with each synapse and thus realizing the classification and regression of input data.

XGBoost and LightGBM

Extreme Gradient Boosting (XGBoost), developed by Chen (*35*), is one of the gradient boosting (GB) algorithms that combining decision tree and boosting method. LightGBM (*36*), another GB algorithm, splits the tree leaf-wise with the best fit while XGBoost split the tree level-wise. The leaf-wise algorithm can reduce more loss than the level-wise algorithm and achieve better accuracy in a short training procedure.

3.3 Data fusion

Since the single sensor positioning is prone to contain measurement errors and loss signals in a complex urban pedestrian networks, we decided to adopt a DF-based location scheme based on the spatial relationship between links to improve the overall accuracy. Refering to Oyama’s link-based measurement model (*1*), we first decompose the time sequence $(1, \dots, t \dots, T)$ into some time periods, and the interval Δt is unchangeable for all time periods. Therefore, the sequence

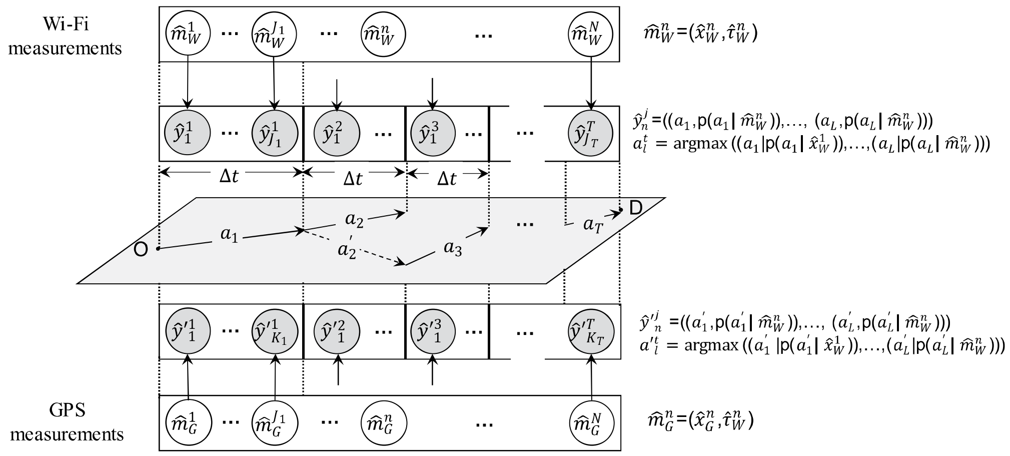

FIGURE 3 A Data Fusion Strategy Incorporating Network Information

of prediction results for a trip $\hat{m}$ is decomposed as $(\hat{y}^1,\ldots,\hat{y}^t,\ldots,\hat{y}^T)$.We choose the data with higher overall accuracy as base data. At each period, use the vector $\hat{y}^t$ predicted from the base data as the initial prediction, and regard the link with the highest probability $p(a_k \mid x_k^t) = max(p(a_1 \mid x_1^t),\ldots,p(a_n \mid x_n^t),\ldots,p(a_N \mid x_N^t))$ as the initial link a_k, and then use the adjacent link set as an indicator to check and replace the initial link. We set connection indicator $\delta(a' \mid a) = 1$ when a and a' are spatially connected, otherwise $\delta(a' \mid a) = 0$. The adjacent link set of link a is $A'(a) = (a'_1,\ldots,a'_C)$, $\delta(a' \mid a) = 1$ for all $a' \in A'$. if an initial link a_τ is not connected to its' upstream or downstream link, and the alternative link a'_τ,we replace a_τ with a'_τ. Figure3 is an example of DF when Wi-Fi data serves as the base data. The solid links on map represents the initial prediction and the dashed link a'_2 represents the replaced prediction.

Generally speaking, link-based measurement models tend to have difficulties in link connection due to their myopic optimization, and the predicted result deviates from the correct path with the error accumulates. In contrast, the fusion scheme we propose can avoid this problem since it is a local optimization strategy. However, it requires a reliable initial prediction and a reasonable time interval Δt.

4. PARAMETER ESTIMATION

4.1 Choice-stage structured network

We construct the route choice model on a choice-stage structured network (CSN) and adopt the definition of CSN in Oyama and Hato's paper (*6*). Let $G = (S,E)$ donate a CSN, where $S = [S_0,\ldots,S_t,\ldots,S_T]$ is the array of state sets which contains state $s_t = (t,a)$, $a \in A$, A is the set of all links of the network; and $E = [E_0,\ldots,E_t,\ldots,E_{T-1}]$ is the array of edge sets with entries of edge $E_t = (s_t,s_{t+1})$, $t \in [0,T]$. Therefore, a route on a CSN can be described as a sequence of states $[s_0,\ldots,s_t,\ldots,s_T]$. Since the state contains both space and time information, state $s_t \neq s_{t'}$, for $t \neq t'$. In other words, an individual has different states at a different time in a CSN, even if the positions in space are the same. This property of CSN can help us to remove cyclic structures from the network, with the spatial cycles remain.

Then we set initial state $s_0 = (0,o)$ and final state $s_T = (T,d)$, o and d are the origin and destination links in A, respectively. T is the time constraint, a key parameter of the CSN, which restrict the stages set S_t through the following mechanism: For any state $s_t = (t, a)$, if an individual can't reach a from o within time t, nor reach d from a within time $T - t$, the state existence indicator $I_t(a) = 0$, otherwise $I_t(a) = 1$. In addition, we set connection indicator $\delta(a' \mid a) = 1$ when a and a' are spatially connected, otherwise $\delta(a' \mid a) = 0$. Thus, the state connection indicator is given as $\Delta(a' \mid a) = I_t(a)\delta(a' \mid a)I_{t+1}(a')$. This procedure can remove unused states, as well as forming a prism to constrain travelers' route choices.

4.2 Recursive logit model

We denote $A(a_j)$ as the set of outgoing links at link $a_j \in A$, i.e., for any link $a \in A(a_j)$, $\delta(a \mid a_j) = 1$. Assume that an individual at state $s_t = (t, a_t)$ always moves to the next state $s_{t+1} = (t+1, a_{j+1})$, $a_{j+1} \in A(a_j)$, which maximizes the sum of instantaneous utility $u(a_{j+1} \mid a_j)$ and the expected downstream utility to destination link d, $V^d(a_{j+1})$. The value function $V^d(a_j)$ can be expressed by using Bellman equation:

$$\begin{aligned} V^d(a_j) &= E\left[\max_{a_{j+1}\in A(a_j)} \{u(a_{j+1} \mid a_j;\theta) + V^d(a_{j+1})\}\right] \\ &= E\left[\max_{a_{j+1}\in A(a_j)} \{v(a_{j+1} \mid a_j;\theta) + V^d(a_{j+1}) + \mu\varepsilon(a_{j+1})\}\right] \end{aligned} \tag{1}$$

Using MNL model, the transition probability from state s_t to state s_{t+1} can be expressed as:

$$p(a_{j+1} \mid a_j) = \frac{e^{\frac{1}{\mu}\{v(a_{j+1}|a_j)+V(a_{j+1})\}}}{\sum_{a'_{j+1}\in A(a_j)} e^{\frac{1}{\mu}\{v(a'_{j+1}|a_j)+V(a_{j+1})\}}} \tag{2}$$

and the route choice probability for route $r = [s_0, \ldots, s_t, \ldots, s_T] = [(0,o), \ldots, (t,a_j), \ldots, (T,d)]$ is:

$$p(r) = \prod_{t=1}^{T-1} p(s_{t+1} \mid s_t) = \prod_{j=1}^{J_n-1} p(s_{t+1} \mid s_t) \tag{3}$$

Since error term ε obeys Gumbel distribution, the value function $V^d(a_j)$ $Eq.(1)$ can also be described as:

$$V^d(a_j) = \begin{cases} \mu \ln \sum_{a_{j+1}\in A} \Delta(a_{j+1} \mid a) e^{\frac{1}{\mu}\{v(a_{j+1}|a_j)+V^d(a_{j+1})\}}, & (t \neq T \cup a \neq d) \\ 0, & (t = T \cup a = d) \end{cases} \tag{4}$$

Because there are no cycles in CSN, we can directly solve the value function from $t = T$, using backward induction.

4.3 Minimum cross-entropy estimation

Maximum likelihood (MLE) is a well-used parameter estimation method for route choice models. However, in this study, we adopt minimum cross-entropy (CE) estimation instead of MLE since the routes inferred by the observation model are in the probabilistic form, consisting of a sequence of vectors of link ID and the corresponding probability. We define the CE function of the vector of parameters θ as follow:

$$CE(\theta) = -\sum_{n=1}^{N}\sum_{j=1}^{J_n-1} p_n(a_j)\ln p_n(a_{j+1} \mid a_j;\theta)$$
$$= -\sum_{n=1}^{N}\sum_{j=1}^{J_n-1} \frac{1}{\mu} p_n(a_j)(v_n(a_{j+1} \mid a_j;\theta) + V_n^d(a_{j+1}) - V_n^d(a_j)) \quad (5)$$

Where N is the number of routes, J_n is the number of links included in route r_n, $p_n(a)$ is the link existence probability inferred from the observation model. $Eq.(5)$ shows that the link transition probability p includes an endogenous variable, and the solution requires computation of the fixed point V^d in $Eq.(1)$. As we discussed in the review section, there are several estimators for dynamic discrete choice models, such as the NFXP algorithm (*30*), the NPL algorithm (*29*), and the MPEC approach (*33*). In this case, we adopt MPEC algorithm.

5. CASE STUDY

5.1 Data collection and network setting

In this study, we use sensor data collected in 2017 on Shibuya station and the surrounding area with a radius of 1 km. The study focused on 40 volunteers, aged 18 to 68, with a 1:1 gender ratio, who lived in the Greater Tokyo Area (Tokyo, Kanagawa, Chiba, and Saitama) and walked through Shibuya station least twice a week. All the volunteers were required to install an application and turn it on for data collection while participating in the following two experiments from March 16th, 2017, to April 12th, 2017.

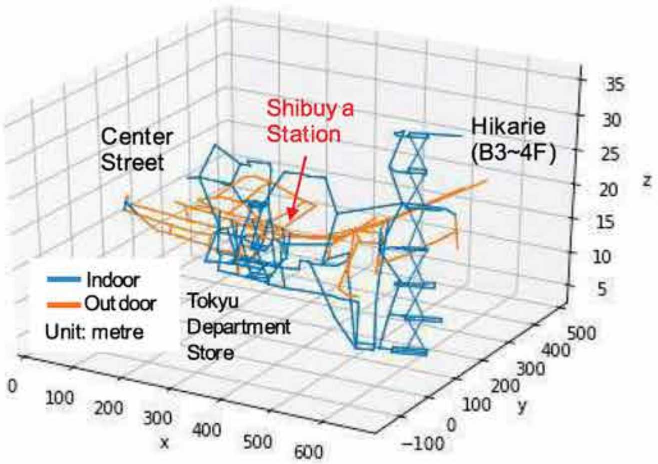

FIGURE 4 Indoor and Outdoor Links in Shibuya 3D Pedestrian Network

The *Experiment*1 asked the volunteers to walk through every single link on the pedestrian network of the survey area, and the recorded data serve as teaching data for location identification. Then, in the *Experiment*2, volunteers can walk freely between the appointed origin and destination within a limited time. The observation result of their route choice behavior is used for pedestrian modeling.

Referring to the GPS coordinates collected in experiment 1, we choose a pedestrian network within a radius of 500 meters around Shibuya Station as the study object. The network covers the roads and passageways within Shibuya station and buildings (Shibuya Mark City, Tokyu Department Store, and Shibuya Hikarie) connected to it, with a maximum of 4 floors above ground and a minimum of 3 floors underground. Although the authentic pedestrian network is composed of directed links, since neither the GPS features nor Wi-Fi features we use for link prediction can reflect the walking direction, we simplify it to an undirected network $\hat{G} = (N, A)$, where N is the set of nodes and A is the set of links.

In building a data set, we assume that the participants always walk at a constant speed and proportion the travel time between two adjacent stamped locations to the links between them according to the link length. This way, we can determine participants' position at each observation time point and label the complete link data with link classes. As a result, we obtained 175 link classes. From the observation data collected in *Experiment*2, we kept the part within the network G, and add 10 more links that might constitute the shortest path to complete the network. Thus, the urban pedestrian network we constructed in this research is $\hat{G}' = (N', A')$, $A' = (a_1, a_2, ..., a_J)$, where $J = 185$. The link distribution is shown in Figure 4.

We screened out outliers in each category(links) by examining velocity and acceleration to reduce inaccuracies and measurement errors in GPS traces. In addition, Smote (*37*), a synthetic minority over-sampling technique, is introduced to reduce sampling bias by generating pseudo instances based on the neighborhood. After screening out the problem data, we obtain a data set for machine learning and observation data for parameter estimation, as shown in Table 1.

TABLE 1 Data Summary

Data Source		*Experiment*1		*Experiment*2	Features
		Training data	Test data	Observation data	
GPS	Indoor	5434	1291	-	3
	Outdoor	5662	1377	-	
	Overall	11096	2668	12190	
Wi-Fi	Indoor	12660	3193	-	8958
	Outdoor	5441	1332	-	
	Overall	18101	4525	33142	

5.2 Prediction result of 3D position observation model

In the training stage, we put training data into learners constructed using machine learning models, including Random Forest, Neural Network, XGBoost, and LightGBM, and did 10-Fold cross-validation for parameter tuning. Since the output of each learner is a probability distribution of the link set, we apply cross-entropy loss as the loss function.

The link with the highest probability is regarded as the final answer and used for calculating model accuracy. The prediction and cross-validation results of each method are shown in Table 2. All the applied classifiers work well on the Wi-Fi data, reaching over 95% test accuracy, while the test accuracy of the GPS data set is basically around 89%. Moreover, the neural network classifier cannot obtain an ideal classification result on GPS data set. That might occur due to the low dimensional data (3 features) and large category (175 classes), which decide the number of nodes in the input and output layers, respectively.

As we discussed in Section 2, based on the positioning principle and data characteristics, both GPS and Wi-Fi positioning have their drawbacks; the former is affected by the building structure and lacks credibility inside the building and when underground; the positioning accuracy of the latter relies on the distribution of access points (APs). Therefore, we examined each models' accuracy indoors and outdoors separately based on the test set labels. According to the result, GPS positioning performs much better outdoors than indoors, and the outdoor accuracy is close to or even slightly higher than that of Wi-Fi positioning in the same environment, which is consistent with the results of the prior study. On the contrary, Wi-Fi positioning is much more stable, with slightly better performance indoors than outdoors, which is reasonable considering that Shibuya, where our study area is located, is one of the busiest commercial districts in the world with a high density of AP distribution.

Figure 5 shows the relationship between the prediction accuracy and the number of detected access points (APs) for each link. The links with low prediction accuracy are mainly distributed in the interval where the number of detected APs is less than 200, both indoor and outdoor, regardless of which model. In other words, Wi-Fi positioning can be vulnerable without a reasonable and dense AP distribution.

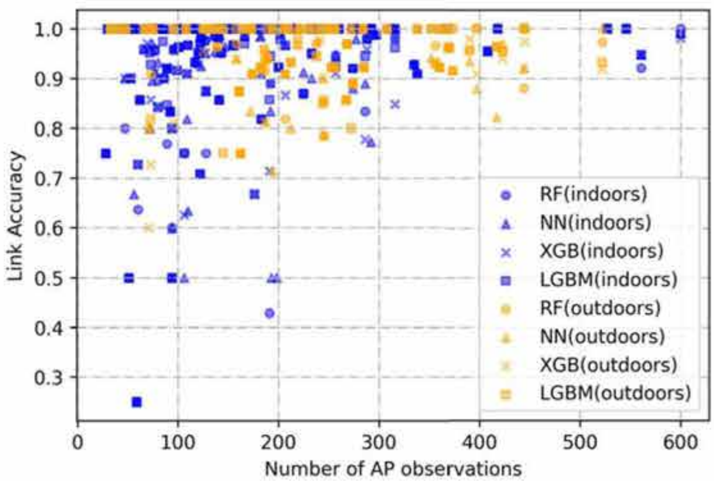

FIGURE 5 The relationship between AP observations and link accuracy

Therefore, we apply the link-based DF strategy shown in section 3.3 for further optimization. Since the GPS positioning is not reliable indoors in this case, we take the result of classifiers trained on Wi-Fi data as the initial result, take a time interval $\Delta t = 10 seconds$, and replace the unconnected outdoor links with the ones with more reasonable position from GPS positioning. As shown in Table 2 , the fusion position model has a better performance than the ones taking GPS or W-Fi data only. In addition, fusion location using multi-sensors is more practical in a complex environment, for that single source of data can be shielded by buildings (GPS data) or lose accuracy in spaces with few AP around (Wi-Fi data).

5.3 Setting of variables and estimation result

In this case study, we define the deterministic component of instantaneous utility function as follows:

$$V(a_{j+1} \mid a_j) = \theta_1 x_{length} + \theta_2 x_{sidewalk} + \theta_3 x_{shop} + \theta_4 x_{indoor} + \theta_5 x_{stair} + \theta_6 x_{escalator} \tag{6}$$

TABLE 2 Prediction result of observation models

		RF		NN		XGBoost		LightGBM	
		Accuracy	CV	Accuracy	CV	Accuracy	CV	Accuracy	CV
GPS	Indoor	83.23%		-		83.87%		83.71%	
	Outdoor	96.46%	86.92%	-	-	95.35%	86.51%	96.66%	87.86%
	Overall	89.37%		-		88.31%		89.12%	
Wi-Fi	Indoor	97.06%		96.33%		96.34%		96.96%	
	Outdoor	96.40%	91.02%	95.80%	89.03%	95.12%	88.85%	96.32%	91.03%
	Overall	96.83%		96.46%		95.96%		96.82%	
Fusion Data		96.88%		-		96.21%		96.97%	

The explanatory variables including link length, sidewalk width, shop distribution rate, indoor dummy, stair dummy and escalator dummy are defined in Table 3.

TABLE 3 Variable Summary

Variable		Explanation
Link Length	x_{length}	Unit:$\times 10 metre$
Sidewalk width	$x_{sidewalk}$	Unit: $metre$
Shop rate	x_{shop}	The shop rate is divided into 5 grades from low to high: 0,1,2,3,4
Indoor dummy	x_{indoor}	1 for indoor link, 0 for outdoor link
Stair dummy	x_{stair}	1 for link having stair on it
Escalator dummy	$x_{escalator}$	1 for link having escalator on it, otherwise 0

Referring to the time limit in the free walk experiment, we set time constraint $T = 20$. We use the results from the Fused data observation models using RF, XGBoost and LightGBM algorithms to do parameter estimation on a combination of the above six variables $\hat{\theta} = (\theta_1, ..., \theta_6)$, Table 4 is the estimation result.

The parameter value for sidewalk width takes a positive number, which reveals that pedestrians tend to walk on streets with wide sidewalks. The negative parameter of the indoor dummy reveals that pedestrians express resistance to the indoor environment while walking. On the other hand, unlike the intuitive judgment that pedestrians tend to choose links that takes less time, the length of links significantly influences pedestrians' route choice. That might be because participants experienced in walking at Shibuya Station make a global decision, focusing on future utility rather than instantaneous utility. To verify this hypothesis, an experiment involving participants lacking Shibuya station knowledge needs to be conducted.

Moreover, the parameter values of shop rate are all positive, and at 1% significance level on observation data processed by XGBoost and LightGBM, which means shops on the links are attractive for the participants and can rise their dwell time. In general, further research is necessary for understanding travelers' route choice in 3D pedestrian network.

TABLE 4 Estimation Result of RL model

	RF		XGBoost		LightGBM	
	parameter	t-value	parameter	t-value	parameter	t-value
x_{length}	0.201	6.28^{**}	0.192	4.81^{**}	0.143	5.76^{**}
$x_{sidewalk}$	0.046	1.35	0.175	4.79^{**}	0.151	3.94^{**}
x_{shop}	0.070	1.14	0.325	6.50^{**}	0.297	4.92^{**}
x_{indoor}	-0.296	-1.70	-0.806	-5.04^{**}	-1.011	-5.80^{**}
x_{stair}	-0.151	-0.16	0.394	0.37	0.381	0.95
$x_{escalator}$	-1.132	-0.92	-1.327	-0.64	-1.010	-0.94
No. of samples		400		400		400
Initial cross-entropy		-97.58		-376.95		-293.07
Final cross-entropy		-69.97		-224.15		-161.87
Cross-entropy ratio		0.28		0.41		0.44

** for 1% significant

6. CONCLUSION

As the complexity of urban networks is increasing dramatically in Japan, Europe, the United States, and China, research into pedestrian dynamics has received widespread attention. Studying the route choice behavior of pedestrians is expected to inform the design of urban pedestrian networks incorporating buildings and pedestrian infrastructure, as well as the implementation of traffic management methods. We proposed a model framework to evaluate and optimally design indoor and outdoor space as an integrated space based on a behavioral perspective.

We proposed a model of pedestrian route choice on a 3D urban network using data collected from GPS and Wi-Fi traces. First, we built a 3D observation model using several machine learning methods and attempted fused positioning based on GPS and Wi-Fi data. Then we adjust the standard RL model and replace the likelihood function with the cross-entropy function to fit probabilistic link sets. Finally, we apply the proposed model to Shibuya Station in Tokyo, Japan. The model exhibited an excellent performance on location identification. We also did parameter estimations on several variables on a CSN, proving that the model owns a good fitness on observation data.

Since Wi-Fi and Bluetooth MAC addresses are refreshed every few tens of minutes, the 3D path choice model can be developed into a sequential model, while some user services provide connection information to Wi-Fi and 5G access points in return. The development of multi-sensor pedestrian models in 3D space and the accumulation of data for the optimal design and control of architectural and exterior spaces will continue to advance, and approaches such as using multi-dimensional features such as Twitter and Instagram as explanatory variables may be considered. In this study, we used Wi-Fi data as input variables for machine learning in the case of indoor spaces. Still, our method can also be extended to image data, which we believe has the potential to improve accuracy. On the other hand, some of the spatial design variables did not show significant values. These results may be due to the heterogeneity of behaviors, and the introduction of latent class models may be an issue. Some nonintuitive decision prone of pedestrians on route choice are discovered as well. To better understand pedestrians' route choice behavior, further study on discount factor and a supplementary Experiment on visitor group (the opposite of the commuter, lacking and are more likely to make a short-sighted decision on route choice) should be conducted.

REFERENCES

1. Oyama, Y. and E. Hato, Link-based measurement model to estimate route choice parameters in urban pedestrian networks. *Transportation research part C: emerging technologies*, Vol. 93, 2018, pp. 62–78.
2. Daganzo, C., *Multinomial probit: the theory and its application to demand forecasting*. Elsevier, 2014.
3. Sasaki, T., Theory of traffic assignment through absorbing Markov process. *Transactions of the Japan Society of Civil Engineers*, Vol. 1965, No. 121, 1965, pp. 28–32.
4. Hato, E., Development of behavioral context addressable loggers in the shell for travel-activity analysis. *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 1, 2010, pp. 55–67.
5. Oyama, Y. and E. Hato, Stochastic Assignment in Time-Structured Networks. *Journal of Japan Society of Civil Engineers, Ser. D3 (Infrastructure Planning and Management)*, Vol. 73, No. 4, 2017, pp. 186–200.
6. Oyama, Y. and E. Hato, Prism-based path set restriction for solving Markovian traffic assignment problem. *Transportation Research Part B: Methodological*, Vol. 122, 2019, pp. 528–546.
7. Madigan, R., Y. M. Lee, and N. Merat, Validating a methodology for understanding pedestrian–vehicle interactions: A comparison of video and field observations. *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 81, 2021, pp. 101–114.
8. El Faouzi, N.-E., H. Leung, and A. Kurian, Data fusion in intelligent transportation systems: Progress and challenges–A survey. *Information Fusion*, Vol. 12, No. 1, 2011, pp. 4–10.
9. Shafique, M. A., E. Hato, and H. Yaginuma, Using probe person data for travel mode detection. *Int. J. Comput. Inf. Syst. Control Eng. World Acad. Sci. Eng. Technol*, Vol. 94, 2014, pp. 1501–1505.
10. Severino, J. V. B., A. Zimmer, T. Brandmeier, and R. Z. Freire, Pedestrian recognition using micro Doppler effects of radar signals based on machine learning and multi-objective optimization. *Expert Systems with Applications*, Vol. 136, 2019, pp. 304–315.
11. Oyama, Y. and E. Hato, A discounted recursive logit model for dynamic gridlock network analysis. *Transportation Research Part C: Emerging Technologies*, Vol. 85, 2017, pp. 509–527.
12. Kasemsuppakorn, P. and H. A. Karimi, A pedestrian network construction algorithm based on multiple GPS traces. *Transportation research part C: emerging technologies*, Vol. 26, 2013, pp. 285–300.
13. Danalet, A., B. Farooq, and M. Bierlaire, A Bayesian approach to detect pedestrian destination-sequences from WiFi signatures. *Transportation Research Part C: Emerging Technologies*, Vol. 44, 2014, pp. 146–170.
14. Patire, A. D., M. Wright, B. Prodhomme, and A. M. Bayen, How much GPS data do we need? *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 325–342.
15. Zhuang, Y., Z. Syed, J. Georgy, and N. El-Sheimy, Autonomous smartphone-based WiFi positioning system by using access points localization and crowdsourcing. *Pervasive and Mobile Computing*, Vol. 18, 2015, pp. 118–136.

16. Zhao, J., H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors. *Transportation research part C: emerging technologies*, Vol. 100, 2019, pp. 68–87.
17. van Oijen, T. P., W. Daamen, and S. P. Hoogendoorn, Estimation of a recursive link-based logit model and link flows in a sensor equipped network. *Transportation Research Part B: Methodological*, Vol. 140, 2020, pp. 262–281.
18. Blewitt, G., M. B. Heflin, F. H. Webb, U. J. Lindqwister, and R. P. Malla, Global coordinates with centimeter accuracy in the International Terrestrial Reference Frame using GPS. *Geophysical Research Letters*, Vol. 19, No. 9, 1992, pp. 853–856.
19. Sharp, I. and K. Yu, *Wireless Positioning: Principles and Practice*. Springer, 2019.
20. Krakiwsky, E. J., C. B. Harris, and R. V. Wong, A Kalman filter for integrating dead reckoning, map matching and GPS positioning. In *IEEE PLANS'88., Position Location and Navigation Symposium, Record.'Navigation into the 21st Century'.*, IEEE, 1988, pp. 39–46.
21. Cascetta, E., A. Nuzzolo, F. Russo, and A. Vitetta, A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks. In *Transportation and Traffic Theory. Proceedings of The 13th International Symposium On Transportation And Traffic Theory, Lyon, France, 24-26 July 1996*, 1996.
22. Ben-Akiva, M. and M. Bierlaire, Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, Springer, 1999, pp. 5–33.
23. Frejinger, E. and M. Bierlaire, Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological*, Vol. 41, No. 3, 2007, pp. 363–378.
24. Gao, S., E. Frejinger, and M. Ben-Akiva, Adaptive route choices in risky traffic networks: A prospect theory approach. *Transportation research part C: emerging technologies*, Vol. 18, No. 5, 2010, pp. 727–740.
25. Fosgerau, M., E. Frejinger, and A. Karlstrom, A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological*, Vol. 56, 2013, pp. 70–80.
26. Mai, T., M. Fosgerau, and E. Frejinger, A nested recursive logit model for route choice analysis. *Transportation Research Part B: Methodological*, Vol. 75, 2015, pp. 100–112.
27. Mai, T., A method of integrating correlation structures for a generalized recursive route choice model. *Transportation Research Part B: Methodological*, Vol. 93, 2016, pp. 146–161.
28. Mai, T., F. Bastin, and E. Frejinger, A decomposition method for estimating recursive logit based route choice models. *EURO Journal on Transportation and Logistics*, Vol. 7, No. 3, 2018, pp. 253–275.
29. Aguirregabiria, V. and P. Mira, Swapping the nested fixed point algorithm: A class of estimators for discrete Markov decision models. *Econometrica*, Vol. 70, No. 4, 2002, pp. 1519–1543.
30. Rust, J., Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 1987, pp. 999–1033.

31. Zimmermann, M. and E. Frejinger, A tutorial on recursive models for analyzing and predicting path choice behavior. *EURO Journal on Transportation and Logistics*, Vol. 9, No. 2, 2020, p. 100004.
32. Hotz, V. J. and R. A. Miller, Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, Vol. 60, No. 3, 1993, pp. 497–529.
33. Su, C.-L. and K. L. Judd, Constrained optimization approaches to estimation of structural models. *Econometrica*, Vol. 80, No. 5, 2012, pp. 2213–2230.
34. Breiman, L., Random forests. *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32.
35. Chen, T. and C. Guestrin, Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
36. Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, Vol. 30, 2017, pp. 3146–3154.
37. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, Vol. 16, 2002, pp. 321–357.