Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 6657486, 14 pages
https://doi.org/10.1155/2022/6657486

Research Article

Exploring for Route Preferences of Subway Passengers Using Smart Card and Train Log Data

Eun Hak Lee,[1] **Kyoungtae Kim**,[2] **Seung-Young Kho**,[1,3] **Dong-Kyu Kim**,[1,3] **and Shin-Hyung Cho**[4]

[1] *Institute of Construction and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*
[2] *Future Transport Policy Research Division, Korea Railroad Research Institute, 176, Cheoldobangmulgwan-ro, Uiwang-si, Gyeonggi-do 16106, Republic of Korea*
[3] *Department of Civil and Environmental Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*
[4] *School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA*

Correspondence should be addressed to Shin-Hyung Cho; scho370@gatech.edu

Received 18 December 2020; Revised 12 October 2021; Accepted 10 November 2021; Published 3 January 2022

Academic Editor: Hai Ying Li

As the mode share of the subway in Seoul has increased, the estimation of passenger travel routes has become a crucial issue to identify the congestion sections in the subway network. This paper aims to estimate the travel train of subway passengers in Seoul. The alternative routes are generated based on the train log data. The travel route is then estimated by the empirical cumulative distribution functions (ECDFs) of access time, egress time, and transfer time. The train choice probability is estimated for alternative train combinations and the train combination with the highest probability is assigned to the subway passenger. The estimated result is validated using the transfer gate data which are recorded on private subway lines. The result showed that the accuracy of the estimated travel train is shown to be 95.6%. The choice ratios for no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips are estimated to be 53.9%, 37.7%, 6.5%, 1.5%, and 0.4%, respectively. Regarding the practical application, the passenger kilometers by lines are estimated with the travel route estimation of the whole network. As results of the passenger kilometer calculation, the passenger kilometer of the proposed algorithm is estimated to be 88,314 million passenger kilometer. The proposed algorithm estimates the passenger kilometer about 13% higher than the shortest path algorithm. This result implies that the passengers do not always prefer the shortest path and detour about 13% for their convenience.

1. Introduction

In 2004, the municipal government of Seoul introduced the automatic fare collection (AFC) system. The AFC system makes it possible to analyze the travel behavior of transit passengers. With smart card data obtained from the AFC system, it has much attention to estimate the travel route of passengers on subway networks [1]. Seoul's transit fare system charges passengers based on their travel distance, so it is essential to ascertain the passenger's travel routes [2]. Smart card data of the AFC system provide travel route information of bus trips and transfer trips between the bus and subway networks [3, 4]. The travel routes of the subway passengers, however, are still hard to identify since the smart card data do not provide route information of subway passengers [5]. The card reader of the subway AFC system is installed at the gates of the station, which is outside of the platform. Since the information is only recorded at the station gates that a passenger departs or arrives, thus there is no way to know which route a passenger has traveled. The crucial problem of estimating the travel routes of subway passengers is that there is no information about transfer trips between public subway lines [6]. Only privately owned lines have installed the transfer gates, which are located on the transfer aisle. Travel route information of trips made through the private lines can be identified with the transfer gate data. For the transfer trips of public lines, the travel route information is not provided since there is no transfer gate at the transfer station.

The travel routes of urban railways have traditionally been estimated based on utility maximization or regret minimization models [7, 8]. However, these models could not be valid for several reasons. The train arrival time is not always consistent with the train schedule in a complex urban railway system. Also, passengers might not choose the estimated travel route depending on their tap-in time and train arrival time. Passengers could choose unexpected travel routes with instantaneous decisions. Thus, the traditional models were not always correct in these specific situations, and the advanced method is required to estimate the travel route [9].

Recently, many studies have explored route preference using smart card data [10–14]. For example, Sun et al. [15] estimated the passenger's location with smart card data of the Singapore MRT system. The spatiotemporal density of passengers was estimated, and the trains' trajectories were identified from the move of estimated density. These results were derived from the railway network in which consecutive trains followed the same route without transfers. Similarly, Kusakabe et al. [16] explored the passenger's train choice behavior with smart card data. The route with the longest in-vehicle time was selected as the traveled route rather than the earliest departing or arriving routes. Lee et al. [17] also estimated the express train choice behavior using smart card data. The Gaussian mixture model was used to decompose the travel time distribution into two distributions, i.e., express train and local train. Each passenger was assigned to an express or local train according to a density probability.

Many previous studies have sought to accurately explore passenger's train preferences using smart card data and train log data, i.e., train logs or train schedules [18, 19]. For example, Sun and Xu [20] estimated the egress time, access time, transfer time, and in-vehicle time with the smart card data, train schedules, and complementary manual surveys. With these estimated attributes, the travel time distribution of each route was established, and the passenger preference was explored. Zhou and Xu [13] also estimated the traveled route to assign passenger flow. With the train schedule data, feasible routes were generated, and each passenger was assigned to the route, which had a minimum surplus time. Similarly, Zhu et al. [21] estimated the train choice behavior with real timetables and smart card data. The choice set was generated by the deletion algorithm, and the route choice probability was estimated by Manski's paradigm. Sun and Schonfeld [22] proposed a route choice model using smart card data. The choice set was generated based on the train schedule connection network. The access time, egress time, and transfer time were considered to assign passengers to the generated route. Similarly, Hong et al. [23] also proposed a train choice model with smart card data and train log data. The passengers who have a unique route were defined as reference passengers, and the traveled routes of passengers who have multi-route were estimated by matching the reference passengers.

Although these previous studies attempted to estimate the travel route, some improvements still remained. First, the accuracy of the route estimation needed to be improved using passenger's experienced travel time attributes, i.e., access time, egress time, transfer time, and in-vehicle time. The distribution forms of the travel time attributes are all different by stations and origin-destination (O-D) pairs. Thus, travel time attributes are required to estimate without the distribution assumption. Second, there was a limitation on validating the model performance since passenger's travel route information, such as transfer information, was not recorded on smart card data. Previous studies have proposed many methods to estimate travel routes. However, there is a limit to identifying the accuracy of the method due to the absence of revealed preference data of travel routes. To shed light on these issues, this study proposed a methodology that estimates passenger's travel route (train) using smart card data and train log data. The contributions of this study were presented as follows: (1) the empirical distribution without distribution assumption was developed to estimate the probability of each travel time attribute; (2) model performance was validated with revealed route information (transfer gate) data; and (3) the practical application, such as efficiency evaluation of each subway line, was performed using estimated results of the whole subway passengers in Seoul.

This study estimated the travel route of individual subway passengers using the smart card data and train log data. The alternative routes were generated based on the train log data. The travel route was then estimated by the empirical cumulative distribution functions (ECDFs) of access time, egress time, and transfer time. With the ECDFs of the time attributes, the train choice probability was estimated for alternative train combinations. Among the alternative train combinations, the train combination with the highest probability was assigned to the subway passenger. The smart card data of the private lines were employed to validate the results of the travel train estimation since it had the exact information about the travel route transaction. The proposed algorithm was then applied to estimate the travel train of all subway passengers on the entire subway network in Seoul.

2. Data Description

2.1. Description of the Network (Seoul Metropolitan Area). The subway network in Seoul consists of 11 lines numbering from 1 to 9, Bundang Line, and Shinbundang Line. The subway network has 327 stations, including 127 transfer stations to serve Seoul and its surroundings. Among 11 lines, Line 9 and the Shinbundang Line are owned by private companies. The total number of trips of the subway network in Seoul is 6,313,176 trips per day. The headway of the subway trains is about 6 minutes on average. The minimum and maximum headways are about 2 and 26 minutes, respectively. There is no way to identify the travel route with the public lines. However, private lines have transfer gates at all transfer stations to collect fares. With the data from the transfer gate, it is possible to validate the results of the travel route estimation. Line 9 consists of 30 stations with nine transfer stations, and the Shinbundang Line consists of 12 stations with five transfer stations. The number of trips of Line 9 and Shinbundang Line is 472,436 trips per day. Since the percentage of private trips accounts for about 7.4% of all trips, it is possible to validate the estimation result.

The travel route estimation for the trips traveled private lines was conducted to validate the performance of the proposed algorithm. The process of estimating train choices for the individual passenger was explained with an illustration network that has two alternative routes for the same O-D pair. The travel route for the subway network in Seoul was also estimated to ascertain the practical applicability of the algorithm. The subway network in Seoul is shown in Figure 1.

2.2. Descriptions of the Smart Card Data and Train Log Data. The smart card data store about 20 million trip information per day, including about 7 million subway trips and 12 million bus trips. The smart card data can be obtained from the Korea Transportation Safety Authority (KTSA) and contain 38 data information for each trip. To estimate the train choice, we used smart card data of October 31, 2017. Among the 38 data information, we used 10; card ID, transaction ID, line ID, boarding station ID, alighting station ID, boarding time, alighting time, total travel time, transfer station ID, and transfer time. The data information related to the transfer is provided only from the trips on the two private lines. Thus, it is possible to identify the travel route of passengers who traveled on private lines. The data information of the smart card data are shown in Table 1.

The train log data contain about 175,000 logs of real-time train operation data per day. The train log data can be obtained from the Open Data Portal (data.seoul.go.kr), and it includes the arrival time information of the train at each station. The reliability of the train log data is ensured because it is the actual arrival time of the train. By integrating train log data with the smart card data, it is possible to estimate the passenger travel route. The train log data used in this study are also from October 31, 2017. It contains eight data information, of which seven data information were used: line ID, arrival time, the direction of train, train ID, train type, boarding station ID, and alighting station ID. The data information of train log data is shown in Table 2.

3. Methodology

The proposed train choice algorithm has two main methodologies, i.e., choice set generation algorithm and empirical cumulative distribution functions (ECDFs). The choice set generation algorithm is used to generate the available train combinations for each passenger. The ECDFs methodology is used to estimate the passenger's choice probability for each alternative. The proposed train choice algorithm consists of seven steps using a choice set generation algorithm and ECDFs. The visualized concept of the train choice algorithm and definition of notations are shown in Figure 2 and Table 3, respectively. For a better understanding of the proposed train choice algorithm, the remainder of the methodology section is organized as follows: the concept of choice set generation algorithm and the concept of ECDFs is described in order. Then, the seven steps of the proposed train choice model are explained step by step.

3.1. Choice Set Generation. In this part, we proposed an algorithm to generate alternative train combinations for an individual passenger using the tap-in time and tap-out time of smart card data, and train arrival time of train log data. The alternative train combination connects the passenger's origin and destination stations during his/her travel time. With the proposed algorithm, it is possible to generate all train choice alternatives for each subway passenger.

The choice set generation is performed for each passenger. Thus, alternative train combinations could be different for the passengers even with the same origin to destination (O-D). The proposed algorithm considered all alternative routes using alternative train combinations during the passenger's travel time. choice combinations during the passenger's travel time. The mathematical expression of the algorithm of generating the alternative train combination is shown in equations (1) to (4). Equation (3) is to find all available trains which depart the origin and arrive at the destination stations between the tap-in and tap-out times of an individual passenger. If there is a transfer station, the train choice combination is generated by connecting transferable trains and the available trains. Equation (4) shows the mathematical expression of the alternative train combination set of the trip *i*.

$$N = \{1, 2, 3, \ldots, 363\}, \tag{1}$$

$$\mathbf{p}_i = \left(t_i^{\text{in}}, t_i^{\text{out}}, o_i, d_i\right), \quad o \in N, d \in N, \tag{2}$$

$$\mathbf{r} = \left(\delta, tr_{\text{in}}^1, tr_{\text{out}}^1, tr_{\text{in}}^2, tr_{\text{out}}^2, \ldots, tr_{\text{in}}^k, tr_{\text{out}}^k, \alpha, o, d\right), \quad o \in N, d \in N, \tag{3}$$

$$R_{OD}(\mathbf{p}_i) = \left\{\mathbf{r} \mid \delta \geq t_i^{\text{in}}, \alpha \leq t_i^{\text{out}}, o = o_i, d = d_i\right\}. \tag{4}$$

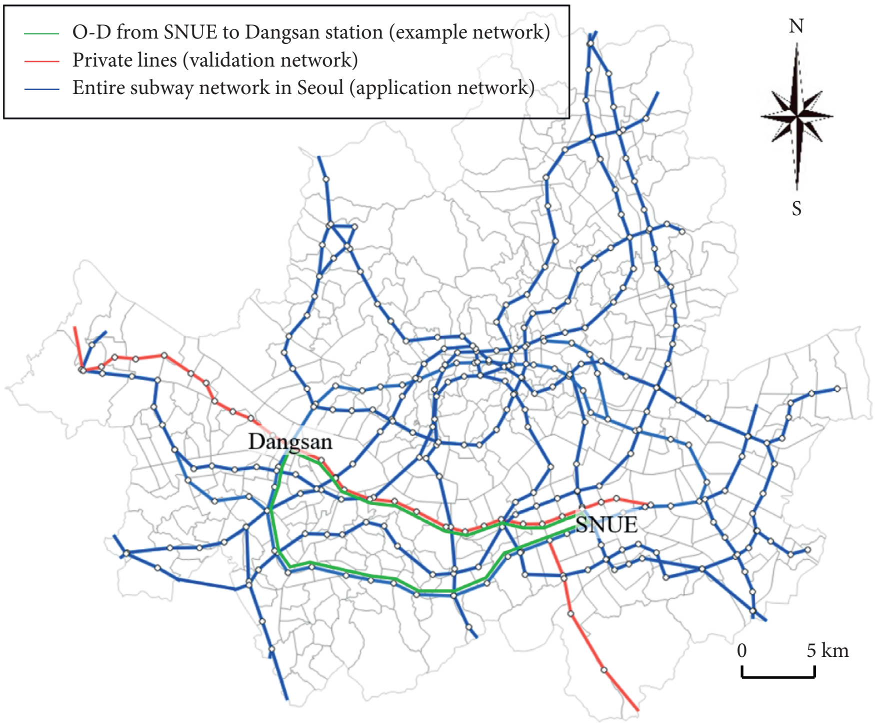

O-D from SNUE to Dangsan station (example network)
Private lines (validation network)
Entire subway network in Seoul (application network)
N
S
Dangsan
SNUE
0
5 km

Figure 1: Subway network in Seoul.

3.2. Empirical Cumulative Distribution Function. The ECDF is a nonparametric estimator of the typical CDF of a random variable. ECDF has an advantage in estimating probabilities because assumptions are relatively free. For example, distributions of the travel time attributes are difficult to define in the specific form since the distribution of each station and O-D pairs is all different. If there are plenty of samples, the ECDF can improve the accuracy of the model. In other words, the ECDF approximates the true CDF with the large samples. It estimates a probability of $1/j$ to each sample, orders the samples from smallest to largest in value, and calculates the sum of the estimated probabilities up to and including each sample value. The result is a step function that increases by $1/j$ at each sample value. The ECDF is usually denoted by f_j or $P_j(X \leq x)$, and mathematical expression is defined as follows:

$$f_j(x) = P_j(X \leq x) = j^{-1} \sum_{i=1}^{j} I(x_i \leq x). \tag{5}$$

$I(x_p \leq x)$ is the indicator function and has two values. If the event inside the brackets occurs, the value is 1, and if not, the value is 0.

$$I(x_p \leq x) = \begin{cases} 1, & \text{when } x_p \leq x, \\ 0, & \text{when } x_p > x. \end{cases} \tag{6}$$

3.3. Train Choice Algorithm. To estimate the passengers' travel train combinations, we developed a train choice algorithm using smart card data and train log data. The proposed algorithm consists of seven steps. Step 1 is to extract information about passengers who have a clear train combination to travel. In this case, the passenger has only one train available to travel from the origin station to the destination station between tap-in time and tap-out time. In Step 2, the time attributes, i.e., access time, egress time, and transfer time, are calculated by the extracted passenger's tap-in time and tap-out time and train arrival time and departure time. In Step 3, the ECDFs of access time, egress time, and transfer time for each station are developed using the calculated time attributes. Step 4 is for generating alternative train choices for a passenger who has more than two alternative trains on his/her route. In Step 5, the choice probability is estimated for each alternative train. The train choice probability is calculated by multiplying the probability of time attribute, i.e., access time, egress time, and transfer time for all of the alternative trains. The probability of each travel time attribute converges to 1 as it approaches the mode value. In step 6, the train combination with the highest choice probability is assigned to a passenger. Step 7 is the iteration step for estimating the next passenger's travel train combination. The mathematical expression of the travel train estimation algorithm is shown in equations (7) to (19).

Step 1. Select the set of passengers who have only one alternative train combination during his/her travel time.

The passenger group with one train available is selected by comparing the tap-in time and tap-out time of smart card data to the train arrival time at the origin station of the train log data. Specifically, all available train combinations during the tap-in time and tap-out time of each passenger are

Table 1: Description of the smart card data.

No.	Data information
1	Card ID*
2	Transaction ID*
3	Mode code
4	Line ID*
5	Name of the transit line
6	Vehicle ID
7	Vehicle number
8	Boarding station ID*
9	Alighting station ID*
10	Name of boarding station
11	Name of alighting station
12	Boarding (tap-in) time*
13	Alighting (tap-out) time*
14	Number of transfer
15	Total travel distance
16	Total travel time*
17	Boarding fare
18	Alighting fare
19	The number of users
20	Boarding violation penalty
21	Alighting violation penalty
22	General user code
23	Student user code
24	Child user code
25	Other user code
26	User division
27	User group
28	Company code
29	Company name
30	Time code
31	Starting run time
32	Ending run time
33	Boarding date
34	Alighting date
35	Year
36	Zone code
37	Transfer station ID
38	Transfer time

*Used in this study.

checked, and a passenger who has only one available train is selected in this step.

$$U = \{\mathbf{p}_i | \mathrm{n}(R_{OD}(\mathbf{p}_i)) = 1, \quad \mathbf{p}_i \in P = \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_i \ldots, \mathbf{p}_{6,313,176}\}\}. \tag{7}$$

Step 2. Calculate the travel time attributes of the set of passengers who have only one alternative train combination.

The access time, the egress time, and the transfer time of individual passengers are estimated using the tap-in time and tap-out time from the smart card data and train arrival time at the origin, transfer, and destination stations.

$$a_i = \delta - t_i^{\mathrm{in}}, \tag{8}$$

$$e_i = t_i^{\mathrm{out}} - \alpha, \tag{9}$$

Table 2: Description of the train log data.

No.	Data information
1	Name of affiliate
2	Line ID
3	Arrival time
4	The direction of the train
5	Train ID
6	Train type
7	Boarding station ID
8	Alighting station ID

$$tr_i = tr_{\mathrm{out}}^k - tr_{\mathrm{in}}^k. \tag{10}$$

Subject to

$$r = \left(\delta,\, tr_{\mathrm{in}}^1, tr_{\mathrm{out}}^1,\, tr_{\mathrm{in}}^2, tr_{\mathrm{out}}^2,\, \ldots, tr_{\mathrm{in}}^k, tr_{\mathrm{out}}^k, \alpha, o, d\right) \in R_{OD}(\mathbf{p}_i). \tag{11}$$

$$\mathbf{p}_i = \left(t_i^{\mathrm{in}},\, t_i^{\mathrm{out}},\, o_i,\, d_i\right) \in U. \tag{12}$$

Step 3. Develop the empirical cumulative distribution function (ECDF) of time attributes.

ECDFs are set up using the access time, the egress time, and the transfer time of individual passengers who have only one train available.

$$F_a^o(a_u^o) = f_j(a_u^o), \quad \text{for } u \text{ s.t. } \mathbf{p}_u \in U, \tag{13}$$

$$F_e^d\left(e_u^d\right) = f_j\left(e_u^d\right), \quad \text{for } u \text{ s.t. } \mathbf{p}_u \in U, \tag{14}$$

$$F_{tr}^k\left(tr_u^k\right) = f_j\left(tr_u^k\right), \quad \text{for } u \text{ s.t. } \mathbf{p}_u \in U. \tag{15}$$

Step 4. Generate alternative train combinations for a passenger who has multiple alternatives.

The set of passengers could be generated when they have multiple trains available at origin, transfer, and destination stations between their tap-in time and tap-out time.

$$M = \{\mathbf{p}_i | \mathrm{n}(R_{OD}(\mathbf{p}_i)) > 1, \quad \mathbf{p}_i \in P = \{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots, \mathbf{p}_i \ldots, \mathbf{p}_{6,313,176}\}\}. \tag{16}$$

Step 5. Calculate the choice probability of each alternative train.

The choice probability of each alternative train was estimated by multiplying three probabilities of access time, transfer time, and egress time. The probability of the mode value was assumed to be 100% since the travel time attributes formed the skewed distribution. As the travel time attributes become closer to the mode value, there will get a higher chance to board the train. Therefore, the probability was defined based on the distance from the mode value as the probability of the corresponding time attributes.

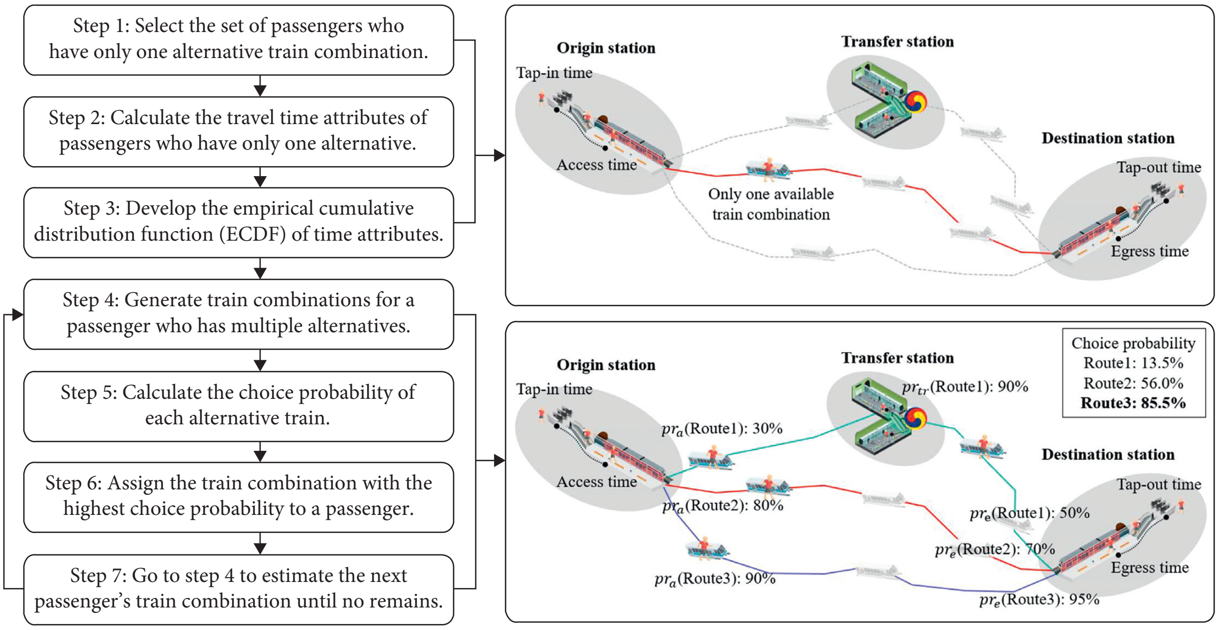

Step 1: Select the set of passengers who have only one alternative train combination.
Step 2: Calculate the travel time attributes of passengers who have only one alternative.
Step 3: Develop the empirical cumulative distribution function (ECDF) of time attributes.
Step 4: Generate train combinations for a passenger who has multiple alternatives.
Step 5: Calculate the choice probability of each alternative train.
Step 6: Assign the train combination with the highest choice probability to a passenger.
Step 7: Go to step 4 to estimate the next passenger's train combination until no remains.
Origin station
Tap-in time
Access time
Transfer station
Only one available train combination
Destination station
Tap-out time
Egress time
Origin station
Tap-in time
Access time
Transfer station
pr_{tr}(Route1): 90%
Choice probability
Route1: 13.5%
Route2: 56.0%
Route3: 85.5%
pr_a(Route1): 30%
pr_a(Route2): 80%
pr_a(Route3): 90%
Destination station
Tap-out time
pr_e(Route1): 50%
pr_e(Route2): 70%
Egress time
pr_e(Route3): 95%

Figure 2: Visualized concept of train choice algorithm.

$$pr = pr_a * pr_e * pr_{tr}, \tag{17}$$

$$pr_a = 1 - \left| F_a^o\left(a_m^o\right) - F_a^o\left(ma_u^o\right) \right|, \tag{18}$$

$$pr_e = 1 - \left| F_e^d\left(e_m^d\right) - F_e^d\left(me_u^d\right) \right|, \tag{19}$$

$$pr_{tr} = 1 - \left| F_{tr}^k\left(tr_m^k\right) - F_{tr}^k\left(mtr_u^k\right) \right|, \quad \text{for } u \text{ s.t. } \mathbf{p_u} \in U \text{ for } m \text{ s.t. } \mathbf{p}_m \in M. \tag{20}$$

Step 6. Assign the train combination with the highest choice probability to a passenger.

Among the multiple train combinations, the train combination with the highest choice probability is assigned to a passenger. The train choice probability is estimated by multiplying the probability of each travel time attribute. The calculation is based on the multiplication rule probability. If the passenger has an alternative route with transfers, the choice probability of transfer is multiplied as a transfer penalty. If not, the train choice probability is estimated with the choice probability of access time and egress time. The mathematical expression of estimating the train choice probability is shown in the following equation:

$$\begin{aligned} & v^* = v, \\ & s.t. pr^v = \max\left(pr^1, pr^2, pr^3, \dots pr^v, \ \dots, pr^w\right). \end{aligned} \tag{21}$$

Step 7. Go to Step 4 to estimate the next passenger's travel train combination until no remains.

The steps from 4 to 7 operate iteratively until estimating all passengers' train choices, since the proposed algorithm estimates the train choice for each passenger.

3.4. Performance Measure for Validating Train Choice.

The performance measures, e.g., precision, recall, accuracy, and F1 score, were used to validate the model performance. The precision, recall, accuracy, and F1 score are well-known measures for validating the performance of the model in each passenger. The values of performance measures were estimated by comparing the passenger's explored route from the assigned train combination and the actual route recorded in smart card data. Precision is defined as the accuracy of estimating true positives from the true negatives and false positives, as in equation (22). The recall is the number of true positives among the true negatives and false positives as in equation (23). The accuracy is the number of true positives and true negatives among all the passengers, as in equation (24). The F1 score is the trade-off between recall and precision, and has equal importance as in equation (25):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{22}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{23}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \tag{24}$$

Table 3: Definition of notations.

Choice set generation algorithm

N: the set of the subway station number

$\mathbf{p}_i$: the vector of the travel attributes of the passenger i

t_i^{in}: the tap-in time of the passenger i

t_i^{in}: the tap-out time of the passenger i

o_i: the origin station of the passenger i

d_i: the destination station of the passenger i

$\mathbf{r}$: the vector of the attributes of the train combination

δ: the train departure time at the origin station

tr_{in}^k: the arrival time of the train for the previous segment (before transfer) at the transfer station k

tr_{out}^k: the departure time of the train for the next segment (after transfer) at the transfer station k

α: the train arrival time at the destination station

o: the origin station of the train combination

d: the destination station of the train combination

$R_{OD}(\mathbf{p}_i)$: the set of the alternative train combination for the passenger i

ECDF

$f_j(x)$: ECDF of the attribute x

Train choice algorithm

Choice set-related notations

U: the set of the passengers who have only one alternative train combination

M: the set of the passengers who have more than two alternative train combinations

P: the set of the passengers

$n(R_{OD}(\mathbf{p}_i))$: the number of the alternative train combination of passenger i

Travel time attribute-related notations

a_i: the access time of passenger i

e_i: the egress time of passenger i

tr_i: the transfer time of passenger i

a_i^o: the access time at the origin station o

e_i^d: the egress time at the destination station d

tr_i^k: the transfer time at the transfer station k

ma_u^o: the mode value of the access time of the passenger u

me_u^d: the mode value of the egress time of the passenger u

mtr_u^k: the mode value of the transfer time at the transfer station k of the passenger u

ECDF-related notations

F_a^o: the ECDF of the access time at the origin station o

F_e^d: the ECDF of the egress time at the destination station d

F_{tr}^k: the ECDF of the access time at the origin station k

Choice probability-related notations

pr: the choice probability of the train combination

pr_a: the probability of access time of the alternative train combination

pr_e: the probability of egress time of the alternative train combination

pr_{tr}: the probability of transfer time of the alternative train combination

v^*: the number of the train combination with the highest choice probability

w: the number of the alternative train combination

pr^v: the choice probability of alternative train combination v

$$F1\ \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{25}$$

where TP is the true positives, FP is the false positives, TN is the true negatives, and FN is the false negatives.

4. Application

4.1. Validation of the Travel Route Estimation Results. The results of estimated travel routes and train combinations for individual passengers are validated with smart card data obtained from two private lines, i.e., Line 9 and the Shinbundang Line. The route information of passengers who get in or get off the private lines as part of their travel routes could be easily produced since the private lines facilitate transfer gates at their transfer stations. The results of the travel route estimation are compared with the actual route of trips recorded in smart card data. For example, O-D pair in Figure 3 was selected to illustrate the process of the train choice estimation. Figure 3 shows the route of the Seoul National University of Education (SNUE) Station to Dangsan Station. There are two alternative routes between SNEU Station and Dangsan Station: no-transfer route and one-transfer route. Route 1 directly connects O-D stations with no transfers, and route 2 contains one transfer at Express Terminal Station on their route. Route 1 is the no-transfer route, which is on a single line. Route 2 is a one-

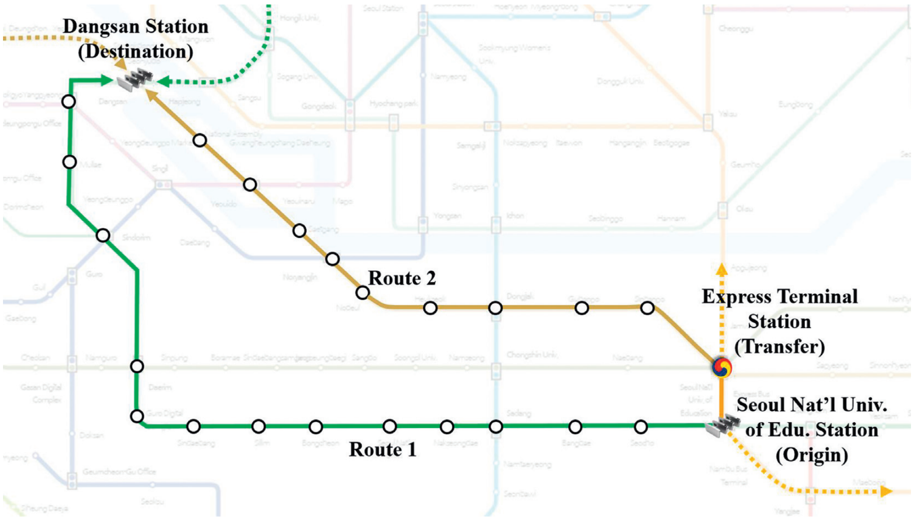


Figure 3: Illustration network with alternative routes from SNUE Station to Dangsan Station.

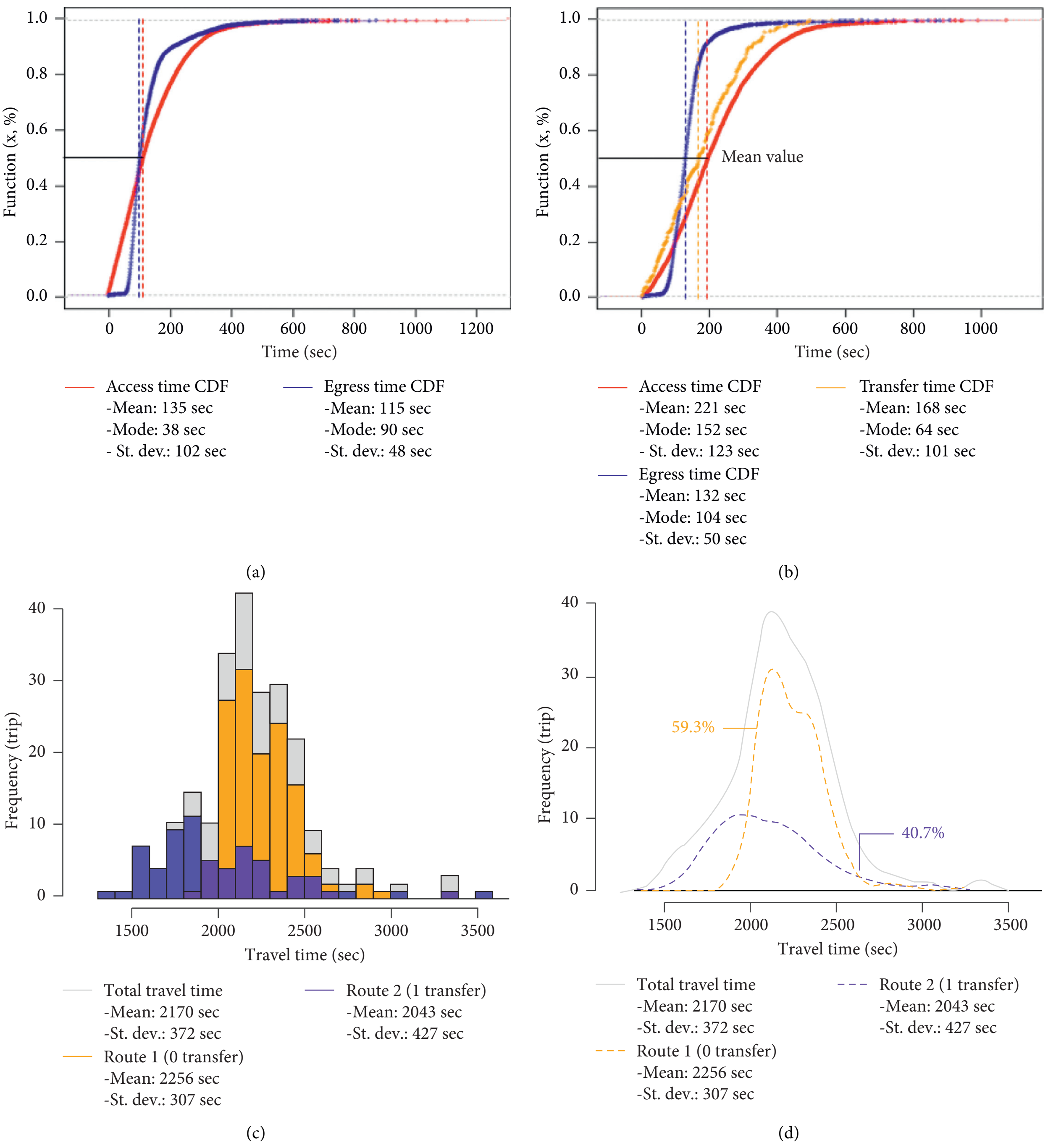
transfer route, where the Express Terminal Station connects the two lines. All ECDFs for each direction of origin station, destination station, and transfer stations were used to select the appropriate travel train combination. The alternative routes from SNUE Station to Dangsan Station are shown in Figure 3.

Figures 4(a) and 4(b) illustrate the cumulative distribution of travel time attributes, which are access time, egress time, and transfer time of routes 1 and 2.

As a result of the developed distributions, the mean of the access time of route 1 was estimated to be 135 seconds. The mode of egress time of route 1 was also estimated to be 38 seconds, and the standard deviation was 102 seconds. The mean, mode, and standard deviation of the egress time of route 1 were estimated to be 115, 90, and 48 seconds, respectively. For route 2, the average of access time, egress time, and transfer time was estimated to be 221, 132, and 168 seconds, respectively. The mode value of access time, egress time, and transfer time of route 2 was estimated to be 152, 104, and 64 seconds, respectively. The standard deviations of access time, egress time, and transfer time were estimated to be 123, 50, and 101 seconds, respectively. Figures 4(c) and 4(d) show the travel time distributions of the two routes. The grey histogram in Figure 4(c) and the grey line in Figure 4(d) represent the total travel time distribution of passengers from SNUE Station to Dangsan Station. This total travel time distribution is shown as the mixed distribution of two routes' travel time. With the distributions of access time, egress time, and transfer time, the total travel time distribution was decomposed by two distributions of respective routes. The results of the decomposed distributions are colored yellow for route 1 and blue for route 2. The mean of total travel time of OD is 2,170 seconds, and the standard deviation is 372 seconds. For route 1, the average travel time is estimated to be 2,256 seconds and the standard deviation is 307 seconds. Route 2 has 2,043 seconds for the average travel time and 427 seconds for the standard deviation of travel time. The result of the travel route estimation from SNUE Station to Dangsan Station is shown in Figure 4.

The comparison analysis was conducted to evaluate the performance of the proposed model. Three comparison models were used to compare with the proposed model. Three comparison models consist of the Gaussian mixture model (GMM) [17], maximum route length model (MRL) [9], and parametric distribution model (PDM) [20]. GMM decomposed the travel time distribution into the number of routes, assuming the Gaussian distribution. GMM assigned the train combination to a passenger with the probability distribution of each route travel time. MRL assigned the train combination to a passenger with the maximum route length (time duration) that fits within the tap-in and tap-out time of the journey. PDM assigned the train combination to a passenger based on the travel time attribute distributions, e.g., access, egress, transfer, and in-vehicle time. The access, egress, and transfer time were assumed to be gamma distribution. The waiting time and in-vehicle time were assumed to be the Poisson and uniform distributions, respectively. Each parameter of distribution was estimated to explore the passengers' route choice preference. Overall, four models, including the proposed model, were compared to evaluate the model performance.

As a result of the comparison analysis, the choice probability of route 1 was estimated to be 54.4% to 64.8%. Among the four models, the proposed model had the most similar probability at 59.3% compared with the actual route choice probability. Regarding individual train combination choice, the F1 scores of GMM, MRL, PDM, and proposed model were estimated to be 0.688, 0.739, 0.918, and 0.963, respectively. Overall, the proposed model showed the highest performance in both aggregated probabilities, such as choice probability and individual choice estimation. PDM also showed good performance with 0.918 F1 score. However, the F1 score of PDM was estimated to be lower than that of the proposed model since the errors due to the assumption of



(a)

(b)

(c)

(d)

Figure 4: Estimation results of SNUE Station to Dangsan Station trips. (a) Cumulative distribution for no-transfer route. (b) Cumulative distributions for one-transfer route. (c) Histogram of travel time. (d) Distribution of travel time.

distribution are involved. Especially, the assumption of uniform distribution had the greatest influence on the inaccuracy. These results implied that the proposed model estimates passengers' train choice preference more accurately than the GMM, MRL, and PDM. The travel route estimation result of the comparison models is shown in Table 4.

The results of the proposed algorithm are validated using the trips made through the private lines. As mentioned before, smart card data from the private lines provide transfer information and make it possible to identify the passenger's travel route.

From smart card data, the number of trips on private lines was counted as 472,436 trips per day. The numbers of no-transfer, one-transfer, two-transfer, and three-transfer trips are counted as 220,239, 241,114, 10,738, and 345, respectively. Table 5 shows the validation results of the travel route estimation of the proposed algorithm compared with the counted number of passengers who get in or get out of the private lines, Line 9 and Shinbundang Line, during their journey. The results

Table 4: Travel route estimation result of the comparison models.

Division	Estimated number of trips		Estimated choice probability (%)		F1 score
	Route 1	Route 2	Route 1	Route 2	
Actual	145	108	57.3	42.7	—
GMM	164	89	64.8	35.2	0.688
MRL	138	115	54.5	45.5	0.739
PA	157	96	62.1	37.9	0.918
Proposed	150	103	59.3	40.7	0.963

GMM: Gaussian mixture model. MRL: maximum route length model. PA: parametric distribution model

of no-transfer trips estimated by the proposed algorithm showed 99.7% of accuracy. For the one-transfer trips, 223,117 trips of 241,114 trips were estimated correctly, and the accuracy was estimated to be 92.5%. As a result of the two- and three-transfer trips, the accuracy was declined to be 81.1% and 71.6%, respectively. Taken together, the accuracy of the estimation result for the total trips was estimated to be 95.6%. Since the number of no-transfer and one-transfer trips accounts for 97.6% of the total validation trip samples, the estimation accuracy of the trips was estimated to be high enough to apply the proposed algorithm to the Seoul subway networks. The result of the travel route estimation is shown in Table 5.

Table 5: Result of travel route estimation with private subway lines.

Division	Precision	Recall	Accuracy	F1 score
No-transfer trips	0.997	1.000	0.997	0.998
One-transfer trips	0.947	0.962	0.925	0.954
Two-transfer trips	0.832	0.946	0.811	0.885
Three-transfer trips	0.789	0.833	0.716	0.811
Total	0.968	0.979	0.956	0.974

4.2. Travel Route Estimation for Subway Network in Seoul. The travel trains for 6,313,176 daily trips were estimated to identify the route choice preference using the proposed algorithm. As results, the numbers of no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips were estimated to be 3,402,763; 2,382,288; 411,475; 91,554; and 25,096 trips, respectively. Regarding the trip ratios of total trips, no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips were estimated to be 53.9%, 37.7%, 6.5%, 1.5%, and 0.4%, respectively. The trip ratios of peak and nonpeak hours show similar patterns. The results of the travel route estimation on the whole network in Seoul are shown in Table 6 and Figure 5.

4.3. Evaluating the Efficiency of Subway Lines in Seoul Using the Proposed Algorithm. The proposed algorithm was applied to evaluate the efficiency of 11 subway lines on the Seoul subway network. The algorithm can produce the passenger kilometer metric for evaluating the transport efficiency of 11 lines. The Seoul Transportation Corporation (STC) has been trying to aggregate link trips using smart card data since those are the basic statistics to operate the subway network. STC roughly calculated the passenger kilometer by assigning the passenger to the shortest path because smart card data do not provide travel route information. Regarding this practical need, the travel route estimation could provide useful statistics such as passenger kilometer. The results of the travel route estimation in this study were used to measure the passenger kilometer of 11 subway lines in Seoul.

The most widely used metric to measure transport efficiency is the value of passenger kilometer [24, 25]. Passenger kilometer is calculated by multiplying the number of passengers by the travel distance. The mathematical expression of the passenger kilometer is shown in the following equation:

$$\mathrm{pkm} = \sum_{g}^{G} \mathrm{tpc}_g \times \mathrm{tdc}_g, \tag{26}$$

where pkm is the passenger kilometer value, i is the travel route ($G = 1, 2, 3, \ldots, g$), tpc is the number of passengers who traveled with the route g, and tdc is the distance of the route g (km).

As a result of the passenger kilometer analysis, the passenger kilometer of STC was estimated to be 78,194 million passenger kilometer, and the passenger kilometer of the proposed algorithm was estimated to be 88,314 million passenger kilometer. Since the STC assigned the passenger to the shortest path, the passenger kilometer of the proposed algorithm was estimated to be about 13% higher than that of STC.

The passenger kilometer and the number of passengers were calculated by 11 subway lines. The result of the passenger kilometer of Line 2 was estimated to be 27,002 million passenger km, which is the highest value among the 11 lines. The lowest value was 1,553 million passenger kilometer, of Line 8. Since Line 2 goes through the major commercial and business areas of central Seoul, the passenger kilometer of Line 2 was estimated to be the highest among the 11 lines. For Line 8, the passenger kilometer was estimated to be the lowest because there are only 16 stations along the line and Line 8 serves on the outskirts of Seoul.

Regarding the passenger kilometer per service distance, the efficiencies of 11 lines are evaluated in the order of Line 2, Line 3, and Line 7. The efficiency order based on the number of passengers per service distance is somewhat different from that of the passenger kilometer unit. The efficiency of 11 lines based on the number of passenger units is evaluated in the order of Line 2, Line 5, and Line 7. The evaluation results of 11 lines based on two metrics are presented in Table 7.

Table 6: Results of travel route estimation for urban subway network in Seoul.

Division	Total trips (trip ratio, %)	Peak hour trips (trip ratio, %)		Nonpeak hour trips (trip ratio, %)
		AM (7:00~9:00)	PM (18:00~20:00)	
No-transfer trips	3,402,763 (53.9)	563,952 (54.1)	513,662 (55.6)	2,325,149 (53.5)
One-transfer trips	2,382,288 (37.7)	386,933 (37.1)	337,247 (36.5)	1,658,108 (38.2)
Two-transfer trips	411,475 (6.5)	70,884 (6.8)	57,512 (6.2)	283,079 (6.5)
Three-transfer trips	91,554 (1.5)	15,753 (1.5)	12,557 (1.4)	63,244 (1.5)
Four-transfer trips	25,096 (0.4)	5,189 (0.5)	3,305 (0.4)	16,602 (0.4)
Total	6,313,176 (100.0)	1,042,711 (100.0)	924,283 (100.0)	4,346,182 (100.0)

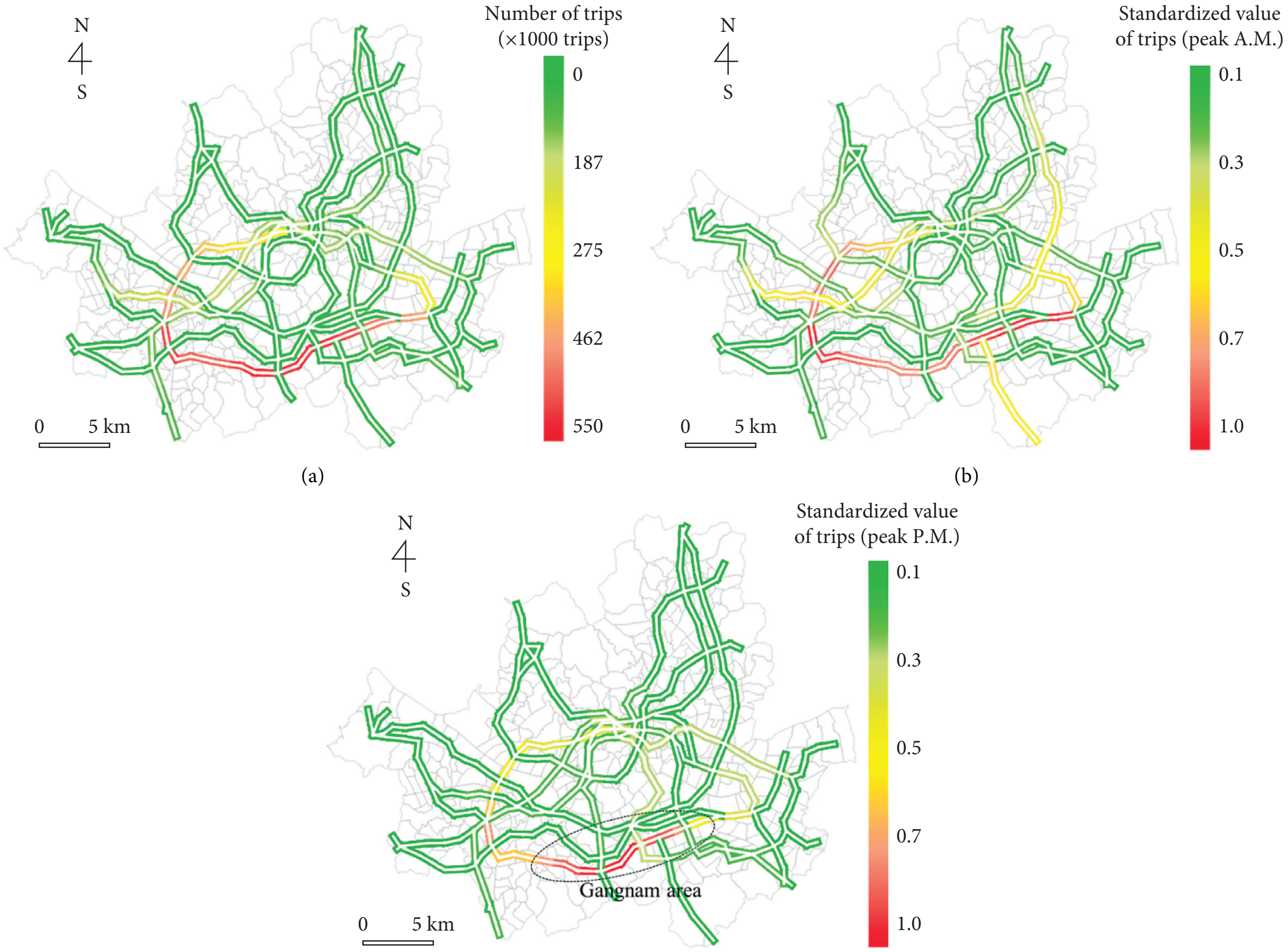


Figure 5: Visualization of estimated link trips of subway network in Seoul. (a) The number of link trips for a day. (b) Link trip density at peak A.M. (c) Link trip density at peak P.M.

Table 7: Results of passenger kilometer for subway lines in Seoul.

Subway lines	Service distance (km)		Number of passengers (trips)			Passenger kilometer (million km)		
	Distance (A)	Rank	Total (B)	Trips/service dist. (B/A)	Rank	Total (C)	Passenger kilometer/service dist. (C/A)	Rank
Line 1	195	1	7,754,053	39,764	9	12,237	63	10
Line 2	**57**	**3**	23,686,939	415,560	1	27,002	474	1
Line 3	55	5	8,027,244	145,950	4	9,523	173	2
Line 4	68	2	5,783,969	85,058	7	6,813	100	8
Line 5	50	8	7,989,509	159,790	**2**	8,105	162	4
Line 6	34	9	4,269,248	125,566	5	3,780	111	5
Line 7	56	4	8,714,031	155,608	3	9,531	170	3
Line 8	17	11	1,477,962	86,939	6	1,553	91	9
Line 9	51	7	3,650,051	71,570	8	5,152	101	7
Bundang Line	53	6	1,451,587	50,055	11	2,900	100	11
Shinbundang Line	31	10	1,015,097	35,003	10	3,168	109	6

Total trips of the Seoul subway network: 6,313,176 trips/day. Estimated passenger kilometer of Seoul network: STC, 78,194 m-pkm (100%); proposed algorithm, 88,314 m-pkm (113%).

5. Conclusion

This study proposed the travel route estimation algorithm using smart card data and train log data. The process of travel route estimation consisted of three stages: (1) generation of the train choice combinations, (2) calculation of passenger travel time attributes, and (3) development of ECDFs. The algorithm was proposed to estimate train choice for an individual subway passenger. The alternative train choice combination was generated using the passenger tap-in time and tap-out time of smart card data, and train arrival time of train log data. The travel time attributes of the passenger were calculated by each alternative train combination. The ECDFs of each type of travel time, i.e., access time, egress time, transfer time, were developed with the trip information that could only be traveled by a single train set. These developed ECDFs were used to estimate the travel route for passengers who have several alternative train combinations. The travel route was deduced by an estimated train combination with the highest probability among the alternative train combinations. The analysis is performed in two stages, i.e., validation with private subway lines and application to the entire subway network in Seoul. For the first stage, the smart card data of the private subway lines were employed to validate the results of the estimated travel train combination, since it has the exact information about the travel route transaction. For the second stage, the proposed algorithm is then applied to estimate the travel train combinations of all subway passengers on the entire subway network in Seoul.

As a result of the comparison analysis, the F1 scores of GMM, MRL, PA, and proposed model were estimated to be 0.688, 0.739, 0.918, and 0.963, respectively. This result implied that the proposed model based on ECDF estimated passengers' choice behavior more accurately than the parametric, nonparametric, and rule-based models. In particular, the proposed model could have strengths in complex subway networks such as many lines, stations, and short headways. As a result of the validation, the accuracy for the no-transfer trips, one-transfer trips, two-transfer trips, and three-transfer trips is estimated to be 99.7%, 95.1%, 84.2%, and 71.2%, respectively. The result of total trips is about 96.9%, which is reasonable to analyze the whole subway network. As a result of the travel route estimation of the whole network in Seoul, the trip ratio for no-transfer, one-transfer, two-transfer, three-transfer, and four-transfer trips was estimated to be 53.9%, 37.7%, 6.5%, 1.5%, and 0.4%, respectively. Regarding the practical application, the passenger kilometers by lines were estimated with the travel route estimation of the whole network. As a result of the passenger kilometer calculation, the passenger kilometer of the proposed algorithm was estimated to be 88,314 million passenger kilometer. Since the STC assigned the passenger to the shortest path, the passenger kilometer of the proposed algorithm was estimated to be about 13% higher than that of STC. Among the 11 subway lines, the passenger kilometer of Line 2 showed the highest value of 27,002 million passenger kilometer.

There are three main contributions to this study. First, the empirical distributions of the travel time attributes, i.e., access time, egress time, transfer time, and in-vehicle time, were developed using smart card data and train log data. Specifically, the subway station's walking characteristics were reflected on access time and egress time without assuming a specific distribution form, i.e., the Poisson and uniform distribution. Second, the real data of passengers' travel routes were used to validate the proposed method. This revealed route information (transfer gate) data provided that the proposed method showed notable accuracy in estimating the travel route of subway passengers. Third, the practical application was performed by estimating whole passengers' travel routes. The results of the efficiency evaluation of each subway line implied that passengers do not always prefer the shortest route.

The results of this paper help subway operators manage in-train and route congestion. The results also contribute to an in-depth investigation of route choice behaviors by quantifying the penalty factors on routes: transfer time and distance, access time and distance, waiting time, the number of stairs, and the congestion rate on the platform. Although we estimated the traveled trains and routes using ECDFs of time attributes, some issues remain. First, the impact of crowding and potentially being left behind needs to be considered. Second, it is required to decompose the walking time and the waiting time distribution for the access time and the transfer time. In addition, information on station amenities, such as restrooms and convenience stores, needs to be considered. Hence, our future work will incorporate crowding and facility factors to estimate the travel route of the subway passengers.

Data Availability

The data used in this research were provided by the Trlab Research Program conducted at the Seoul National University, Seoul, Republic of Korea. The data are available when readers ask the authors for academic purposes.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this study.

Authors' Contributions

Eun Hak Lee provided the software, wrote the original draft, investigated the data, visualized the data, and validated the data. Kyoungtae Kim collected data; wrote, reviewed, and edited the manuscript; and acquired funding. Seung-Young-Kho investigated the data, validated the data, and wrote, reviewed, and edited the manuscript. Dong-Kyu Kim conceptualized the data, supervised the data, designed methodology, investigated the data, involved in formal analysis, wrote, reviewed, and edited the manuscript, and acquired funding. Shin-Hyung Cho conceptualized the data, developed the methodology, investigated the data, involved in formal analysis, and wrote, reviewed, and edited the manuscript.

Acknowledgments

This research was supported by a grant from the R&D Program of the Korea Railroad Research Institute, Republic of Korea.

References

[1] M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

[2] E. H. Lee, H. Shin, S.-H. Cho, S.-Y. Kho, and D.-K. Kim, "Evaluating the efficiency of transit-oriented development using network slacks-based data envelopment analysis," *Energies*, vol. 12, no. 19, p. 3609, 2019.

[3] W. Jang, "Travel time and transfer analysis using transit smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2144, no. 1, pp. 142–149, 2010.

[4] J. Y. Park, D. J. Kim, and Y. Lim, "Use of smart card data to define public transit use in Seoul, Republic of Korea," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2063, pp. 3–9, 2008.

[5] Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual mobility prediction using transit smart card data," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 19–34, 2018.

[6] E. H. Lee, H. Lee, S.-Y. Kho, and D.-K. Kim, "Evaluation of transfer efficiency between bus and subway based on data envelopment analysis using smart card data," *KSCE Journal of Civil Engineering*, vol. 23, no. 2, pp. 788–799, 2019.

[7] B. Si, M. Zhong, J. Liu, Z. Gao, and J. Wu, "Development of a transfer-cost-based logit assignment model for the Beijing rail transit network using automated fare collection data," *Journal of Advanced Transportation*, vol. 47, no. 3, pp. 297–318, 2013.

[8] X. Gong, G. Currie, Z. Liu, and X. Guo, "A disaggregate study of urban rail transit feeder transfer penalties including weather effects," *Transportation*, vol. 45, no. 5, pp. 1319–1349, 2018.

[9] E. Van Der Hurk, L. Kroon, G. Maróti, and P. Vervest, "Deduction of passengers' route choices from smart card data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 430–440, 2014.

[10] K. K. A. Chu and R. Chapleau, "Enriching archived smart card transaction data for transit demand modeling," *Transportation Research Record*, vol. 2063, pp. 63–72, 2008.

[11] N. Nassir, M. Hickman, and Z.-L. Ma, "A strategy-based recursive path choice model for public transit smart card data," *Transportation Research Part B: Methodological*, vol. 126, pp. 528–548, 2019.

[12] J. Chan, *Rail Transit OD Matrix Estimation and Journey Time Reliability Metrics Using Automated Fare Data*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2007.

[13] F. Zhou and R.-H. Xu, "Model of passenger flow assignment for Urban rail transit based on entryand exit time constraints," *Transportation Research Record*, vol. 2284, pp. 57–61, 2012.

[14] W. Zhu, H. Hu, and Z. Huang, "Calibrating rail transit assignment models with genetic algorithm and automated fare collection data," *Computer-Aided Civil and Infrastructure Engineering*, vol. 29, no. 7, pp. 518–530, 2014.

[15] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system," in *Proceedings of the . ACM SIGKDD Int. Workshop Urban Comput.*, pp. 142–148, Beijing, China, August 2012.

[16] T. Kusakabe, T. Iryo, and Y. Asakura, "Estimation method for railway passengers' train choice behavior with smart card transaction data," *Transportation*, vol. 37, no. 5, pp. 731–749, 2010.

[17] E. H. Lee, I. Lee, S.-H. Cho, S.-Y. Kho, and D.-K. Kim, "A travel behavior-based skip-stop strategy considering train choice behaviors based on smartcard data," *Sustainability*, vol. 11, no. 10, p. 2791, 2019.

[18] D. Hörcher, D. J. Graham, and R. J. Anderson, "Crowding cost estimation with large scale smart card and vehicle location data," *Transportation Research Part B: Methodological*, vol. 95, pp. 105–125, 2017.

[19] W. Li, Q. Luo, Q. Cai, and X. Zhang, "Using smart card data trimmed by train schedule to analyze metro passenger route choice with synchronous clustering," *Journal of Advanced Transportation*, vol. 2018, Article ID 2710608, 13 pages, 2018.

[20] Y. Sun and R. Xu, "Rail transit travel time reliability and estimation of passenger route choice behavior,"

Transportation Research Record: Journal of the Transportation Research Board, vol. 2275, no. 1, pp. 58–67, 2012.

[21] W. Zhu, W. Wang, and Z. Huang, "Estimating train choices of rail transit passengers with real timetable and automatic fare collection data," *Journal of Advanced Transportation*, vol. 2017, Article ID 5824051, 12 pages, 2017.

[22] Y. Sun and P. M. Schonfeld, "Schedule-based rail transit path-choice estimation using automatic fare collection data," *Journal of Transportation Engineering*, vol. 142, no. 1, Article ID 04015037.

[23] S. P. Hong, Y. H. Min, M. J. Park, K. M. Kim, and S. M. Oh, "Precise estimation of connections of metro passengers from Smart Card data." *Transportation*, vol. 43, pp. 749–769, 2014.

[24] Uic Activity Report 2018, "International Union of Railway," 2018, https://uic.org/.

[25] B. Feng, E. H. Park, H. Huang et al., "Discrete element modeling of full-scale ballasted track dynamic responses from an innovative high-speed rail testing facility," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2673, no. 9, pp. 107–116, 2019.