Received March 18, 2020, accepted March 25, 2020, date of publication April 6, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985734

Analysis of Subway Passenger Flow for a Smarter City: Knowledge Extraction From Seoul Metro's 'Untraceable' Big Data

HYUNKYUNG SHIN
Department of Financial Mathematics, Gachon University, Seongnam 13120, South Korea
e-mail: hyunkyung@gachon.ac.kr

ABSTRACT Timely and efficient analysis of big data collected from various gateways installed in a smart city is an intractable problem and requires immediate priority. Given the stochastic and massive nature of big data, the existing literature often relies on artificial intelligence techniques based on information theory. As a new approach, this paper presents a knowledge extraction method based on an analysis of Seoul Metro's 'untraceable' ridership big data. Without identification information, the untraceable ridership data only shows the hourly accumulation of station entry and exit information. To reconstruct the missing information in the data set, this study proposes a fluid dynamics model and adopts a heuristic genetic algorithm based on optimization theory as the problem solver. The result of our model presents the distribution of the elapsed time defined on an hourly basis taken until a passenger returns to the station they departed from. To validate our model, we acquired subway ridership data with passengers' identification with permission from Seoul Metro. This paper presents two novel aspects of subway ridership, namely the dependency on departure time and the discrepancy between weekend and weekday traffic. Our analytical approach contributes to solving the problem of extracting hidden knowledge from big collection of data missing critical information, e.g., constantly and autonomously gathered data fragments from numerous gateways in smart cities.

INDEX TERMS Inverse problem, genetic algorithm (GA), optimization, wave decomposition, harmony search algorithm, mass conservation law, data mining, outdoor duration time, Seoul metro subway ridership.

I. INTRODUCTION

A subway system is one of the most important infrastructures of a smart city and subway operations generate big data continuously and massively from the numerous sensor-equipped gateways. Most of the existing research regarding subways has focused on analyzing the number of passengers using a particular subway station or subway line. Through this type of analysis, policy data on the facilities and staffing of subway stations or subway lines can be derived. As the concept of the smart city spreads and efforts to implement it increase, there is a growing need for research that goes beyond existing trends [1], [2]. The ultimate goal of smart city implementation is to improve the quality of life of the citizens living in such cities [3]. To this end, as Lytas et al. insist, big data analysis requires a different perspective [4].

Despite the potential value of big data, we often face difficult challenges to extract meaningful patterns from big data. The goal of this paper is to propose an effective analytical approach to the problem by analyzing the 'untraceable' big data collected from Seoul Metro, South Korea. Seoul is the capital of South Korea and is the country's center of politics, economy, culture, and education. It can be stated with certainty that Seoul is a representative mega city. The population of Seoul was 9, 776, 000 as of 2017. Most of the mega cities suffer from traffic congestion, and Seoul is no exception. The resulting traffic congestions are one of the factors that reduce the quality of life of residents living in mega cities. Residents in a mega city may often use the subway to depart and arrive on time. In this regard, the average daily passenger load of Seoul Metro was 7, 793, 000 as of 2017 [5]. Passengers who use the Seoul Metro tag their tickets every time they enter and exit through subway ticket gates. These tags are stored in a boarding database and produce big data.

As an attempt to develop an analytical approach to finding the hidden patterns from big data, we tapped into the Seoul Metro's untraceable big data with permission. The untraceable dataset has no identification information regarding

The associate editor coordinating the review of this manuscript and approving it for publication was Miltiadis Lytras.

passengers who enter and exit the stations. The ridership data only contains passengers' hourly cumulative entry and exit counts. Containing 1.8 billion passenger entry/exit records, the database we accessed contains approximately 2 million lines of entry and exit data during a year's period, from Mar. 2016 to Feb. 2017. Without passenger identification (ID) records, the data contains no information regarding how long it takes for a specific passenger to return back to the station if the passenger uses the same station for entry and exit.

The goal of this paper is to develop an effective analytical approach to extract hidden knowledge from Seoul Metro's untraceable big data. To accomplish this goal, this study adopted a fluid dynamics model, applied an optimization solver, and generated a heat map that predicts the lifestyles of residents who use the subway. Through these analytical approaches, this study attempts to find a way to determine how long a passenger using a particular subway station is active elsewhere before returning to the subway station. If we can determine this, we can predict the lifestyles of users living or working in the area around a specific subway station. For example, when people return after 2 — —2 hours to the subway on a weekday, they are likely 'shopping' or performing 'leisure' activities rather than working at an office. In this manner, by finding a methodology for predicting the lifestyles of subway station users, which is expected to be the main public transportation of smart cities, it is possible to develop policies that can improve the quality of life of smart city citizens.

Toward this purpose, the paper proceeds as follows. The second section reviews related existing studies. The third section presents a theoretical perspective based on a mathematical model that can analyze subway usage data for one year. The fourth section describes an analytical method and presents a software tool that can analyze one-year data based on the defined mathematical model. The fifth section performs the computations to solve the problems. The sixth section presents test results and performs validation. The seventh section presents discussions regarding the analysis results and concludes the paper.

II. RELATED STUDIES

Related research to our work is categorized into six groups: smart city, subway system, daily routine analysis, human activity recognition, unveiling of new associations from large datasets, and heuristic GA-based harmony search.

Margarita [1] conferred a new definition of a smart city. Visvizi and Lytras [2], Lytras and Visvizi [3], Lytras [4], and Lytras *et al.* [10] proposed policy making considerations, provided a roadmap toward the evolution of big data, and raised the issue of "normative bias" problems related to smart city research. Xiaolei *et al.* [11] demonstrated a data-mining procedure using transit data for Beijing, China. Christina and Konstantinos [12] argued in favor of combining optimization models with data originating from information technology services (ITS). Rosario *et al.* [13] reviewed the advancement of emergent technologies, and their implementations and applications with respect to smart power grids and cities. Andrè Luis *et al.* [14] identified the most important drivers for smart cities from the perspectives provided by professionals from four broad areas of expertise: applied social sciences, engineering, exact and Earth sciences, and human sciences. Joshua *et al.* [15] insisted that cities cannot complete the evolution into smart cities on their own without support from national governments. Murad *et al.* [16] proposed energy-aware communication systems for the Internet of Things (IoT) environment. Lee *et al.* [17] proposed a conceptual framework for building smart cities to conduct comparative case studies by integrating six different perspectives.

Yang *et al.* [18] considered the optimization problem for timetables in subway systems. Subway passenger flow forecasting models were presented for peak-hour flow by Pan *et al.* [19], for special event occurrences by Ni *et al.* [20], and for the Beijing subway using spatiotemporal correlations by Wang and Cai [21]. Machine learning methods were adopted for various subway passenger flow prediction models: Wang *et al.* [22] used the radial basis function (RBF) and a support vector machine (SVM), and Sun *et al.* [23] applied wavelet-SVM to the Beijing subway system. A stochastic approach using a Markov chain Monte Carlo (McMC) model for the route-use patterns of metro passengers was presented by Lee and Sohn [24]. A Bayesian approach taken by Sun et al. [25] is noteworthy. Passenger distributions on subway platforms were studied using ant colony optimization by Yang *et al.* [26].

With the prevalent use of displacement sensing devices that include geo-positioning systems (GPS) and smart cards, daily route analysis has attracted attention, particularly in terms of marketing and public transportation policy. For example, Tao *et al.* [27], Liu *et al.* [28], and Flognfeldt [29] presented methods for traffic pattern analysis for the bus, taxi, and tourism industries, respectively, and Li *et al.* [30] presented a morning commute route analysis regarding public transportation. With the growth of research interest, large-scale public transportation datasets have become available to support research activities, including the datasets provided by Hodge [31] and Karg and Kirsch [32].

Compared to macro-level route analysis as in the above-mentioned studies, low-level human activity recognition using mobile sensors has attracted considerable attention. Kazu *et al.* [33] summarized early stage theories and presented an analytic perspective of trip chaining behavior in Osaka, Japan. Padmaja *et al.* [34] examined reality mining with respect to human behavior analysis, whereas Blanke and Schiele [35] suggested a discriminative classifier for daily routine activities such as working and commuting, and Nikola et al. [36] adopted the inference rule to predict causal relationships between situations and actions. Recently, aggregation with ambience sensing toward activities of daily living (ADL), particularly as it relates to nursing homes, has become a major research topic [37]–[41]. A public dataset for this type of study has been made available [42], [43].

The recent trend of using data-mining techniques in combination with machine learning was thoroughly introduced by Witten *et al.* in [44]. Yadav *et al.* [7] adopted algorithms to handle both structured and unstructured big data, and Wu and Theodoratos [6] employed an incremental frequency computation method to extract non-redundant maximal homomorphic patterns. Chen *et al.* [45] applied a data-mining algorithm to IoT to extract hidden information from data, and David *et al.* [46] introduced a mathematical measure, namely, maximal information-based nonparametric exploration statistics, to identify and classify novel relationships in large datasets. Finally, researchers at Princeton University [47] applied data-mining to explore hidden patterns between a knowledge machine (KM) and its performance.

A harmony search (HS) is a stochastic-genetic-algorithm-based technique that we adopted in this study to solve an optimization problem. Its application area was recently expanded by Geem *et al.* [48], Geem [49], Mahdavi *et al.* [50], Omran and Mahdavi [51], and Geem and Sim [52]. Criticism regarding the method was raised by Chen *et al.* [45]. For the past three decades, heuristic genetic algorithms (Gas) have served as advanced tools for large scale optimization problems and specialize in escaping from local extrema [53]–[56]. Recently, GAs have drawn attention in various academic and industrial fields as a new paradigm for convolutional neural network (CNN)-based deep learning techniques because of their hyper-parameter optimization capability [28], [57].

III. THEORETICAL PERSPECTIVE

Extraction of hidden patterns subtly buried in big data has been a central topic in various advanced prediction scientific fields [6]–[8] Our main interest is on how to infer new information by modeling the inverse problem from a large-scale legacy relational database. The inverse problem in mathematical science has three components: observational data (d_{obs}), an underlying natural law (F), and model parameters of interest (P). It finds the model parameters $p \in P$, which produce the observational data d_{obs} :

$$d_{\text{obs}} = F(p), \tag{1}$$

where F and p denote the underlying natural law and model parameters [9], respectively. For example, consider the Earth's gravitational field based on the density of subsurface rocks. Denote this density as d_{obs}, the observational result of local gravity. Newton's law of gravity is well known and is given as

$$d_{\text{obs}} = \frac{G \cdot p}{r^2}, \tag{2}$$

where d_{obs} is a measure of local gravity, G is the universal gravity constant, p is the local mass of the rock in the subsurface, and r is the distance from the mass to the observation point. Consider the case in which measurements are obtained from two locations: d_1 and d_2. We then have two unknown masses: p_1 and p_2. We denote r_{ij} as the distance between the i-th observation point and the j-th mass.

$$\begin{bmatrix} d_{\text{obs}1} \\ d_{\text{obs}2} \end{bmatrix} = \begin{bmatrix} \frac{G}{r_{11}^2} & \frac{G}{r_{12}^2} \\ \frac{G}{r_{21}^2} & \frac{G}{r_{22}^2} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \tag{3}$$

The solution for the linear system defined by Eq.3 is straightforward.

$$\vec{p} = F^{-1}\vec{d}_{\text{obs}} \tag{4}$$

In this manner, the unknown property of the underground rock can be identified by the observed data d_{obs} through the natural law of gravity.

We presented a simple working example for applying the inverse problem to evaluate a parameter of interest. In this study, our relational database table includes seven field columns: date, subway line ID, station ID, station name, hour, hourly cumulative entry counts, and hourly cumulative exit counts. Let us assume that all the passengers take round trips and they must return to the station from which they departed, i.e., depart at station A, exit at station B, after spending several hours elsewhere, enter station B, and exit at station A. The assumption seems unrealistically strong. However, we can validate that it is actually acceptable, which is summarized later, in the (§.VI Results) section using a ground truth dataset. Briefly speaking, over 58% of ridership data are round trips, and approximately 93% of the round trips satisfy the above assumption. From the ridership database, we construct a new database. A unit data structure of the new database is "passenger transit", which is defined as a round trip by a passenger (satisfying the assumption above), which consists of two tuples: times of entry and exit, both at station s.

$$T_p^s(d) := (T_i|s)\ (T_e|s) \tag{5}$$

Here, p, s, d, T_i, and T_e denote a passenger, station ID, date, time of entry, and time of exit, respectively. The ridership data for a specific date and station s, $R^s(d)$, is then expressed as

$$R^s(d) = \sum_p T_p^s(d). \tag{6}$$

Thus, the total ridership data during the years 2016 and 2017 corresponds to

$$R(2016, 2017) = \sum_d \sum_s \sum_p T_p^s(d), \tag{7}$$

where 2016 and 2017 indicate the years. However, a round trip instead of a single trip must be recorded, which is explained later. For the round trip of a passenger, the data structure is a tuple of passenger transit defined as

$$\left(T_p^{S_0}(d)|T_p^{S_1}(d)\right) := \left((T_i|s_0)\ (T_e|s_0)\ |(T_i|s_1)\ (T_e|s_1)\right), \tag{8}$$

where the first element of the tuple is the transit record for traveling from station S_0 to station S_1, and the second element is the reverse. Taking the ridership data of the Seoul

Metro subway, R, as the observational data, our ultimate goal is to reconstruct $T_p(d)$ for each passenger's trip record. However, this requires too many parameters, including the personal identification of each passenger of the Seoul metro subway, which is not feasible. Instead, we delimit our study to reconstruct the average distribution of the outdoor duration time (**ODT**) of passengers in terms of $E\left\langle T_p^s \right\rangle_p$, which is the ensemble average of transit records in terms of passengers. The measure of a passenger's **ODT** is defined by the difference between the exit and entry times of $T_p^{S_0}(d)$. Consider the following episode: a passenger enters station A at the hour h^i_A, is transported to station B, exits station B at the hour h^e_B, engages in personal activities outside the subway station, enters station B at the hour h^i_B, is transported to station A, and exits from station A at the hour h^e_A. The measure of **ODT** is defined as

$$\textbf{ODT} := h^e_A - h^i_A \tag{9}$$

Here, we address the following question: is it feasible to estimate this **ODT** measure from the observational data? The immediate problem is that our observational data do not include a passenger's ID. How can we identify a passenger who has entered station A in the morning and returns back to the same station in the evening without knowing their ID? Our solution to the question is to adopt the inverse problem by applying the conservation law of mass to traffic.

IV. DATA AND MODEL DESCRIPTION

Consider a simple example to understand how to (partially) reconstruct the unavailable information. The Seoul Metro subway system collects transit records of passengers, as shown in Table 1. A passenger with ID-001 enters station A at 07:00 and exits station B at 7:30. After 10 hours pass, the passenger with ID-001 enters station B at 17:00 and exits station A at 17:30. On the other hand, a passenger with ID-002 makes a round trip from station A to station C during the period between 08:00 and 18:30. For the sake of argument, assume that this raw data is inaccessible; alternatively, only ridership data logged at each station (see Table 2) is available.

TABLE 1. List of subway transit records. Each record contains two logs: station name, and time of entry and exit. According to the list, we can infer that the passenger with ID-001 made a round trip between *A* and *B*, and the one with ID-002 made a round trip between *A* and *C*.

PID	Entry Station	Entry Time	Exit Station	Exit Time
ID-001	*A*	07:00	*B*	07:30
ID-002	*A*	08:00	*C*	08:30
ID-001	*B*	17:00	*A*	17:30
ID-002	*C*	18:00	*A*	18:30

At each subway station, hourly cumulative counts of passengers at the entry and the exit are recorded. The transit records shown in Table 1 are reflected as ridership data at the three stations A, B, and C, as shown in Table 2.

TABLE 2. Subway ridership data logged at each station. The raw data is the list of transit records shown in Table 1.

(a) STATION *A*

H	# En	# Ex
07	1	0
08	1	0
09	0	0
17	0	1
18	0	1
19	0	0

(b) STATION *B*

H	# En	# Ex
07	0	1
08	0	0
09	0	0
17	1	0
18	0	0
19	0	0

(c) STATION *C*

H	# En	# Ex
07	0	0
08	0	1
09	0	0
17	0	0
18	1	0
19	0	0

From the ridership data of "STATION A" shown on the left in Table 2, it can be observed that two passengers made round trips from station A, but it is not feasible to identify which passenger returned at 17:00 to station A (did the passenger enter at 07:00 or 08:00), because passenger ID information is unavailable when producing ridership data. There are two possible scenarios for the transit records of the passengers, as presented in Table 3.

TABLE 3. Two possible passenger transit scenarios.

Hour	Total	ELAPSED TIME BEFORE RETURN					
		7h	8h	9h	10h	11h	12h
07:00	1	0	0	0	1	0	0
08:00	1	0	0	0	1	0	0

(a) Scenario 1 for transit records at station *A*

Hour	Total	ELAPSED TIME BEFORE RETURN					
		7h	8h	9h	10h	11h	12h
07:00	1	0	0	0	0	1	0
08:00	1	0	0	1	0	0	0

(b) Scenario 2 for transit records at station *A*

In scenario 1, the passenger departed at 07:00 from station A and returned at 17:00 (after 10 hours) and the second passenger departed at 08:00 returned at 18:00 (also after 10 hours). In scenario 2, the former departed at 07:00 and returned at 18:00 (after 11 hours) and the latter departed at 08:00 and returned at 17:00 (after 9 hours). The two scenarios are equally probable if there is no further information available. Here is where the main idea of the conservation law of mass is applied. The transit records at station B and C (as shown in Table 4) imply that scenario 1 matches 100% while scenario 2 represents a mismatch.

TABLE 4. Transit records at station *B* and *C*. These can be obtained from Table 2.

Hour	Total	ELAPSED TIME BEFORE RETURN					
		10h	11h	12h	13h	14h	15h
17:00	1	0	0	0	0	1	0

(a) For transit records at station *B*

Hour	Total	ELAPSED TIME BEFORE RETURN					
		10h	11h	12h	13h	14h	15h
18:00	1	0	0	0	0	1	0

(b) For transit records at station *C*

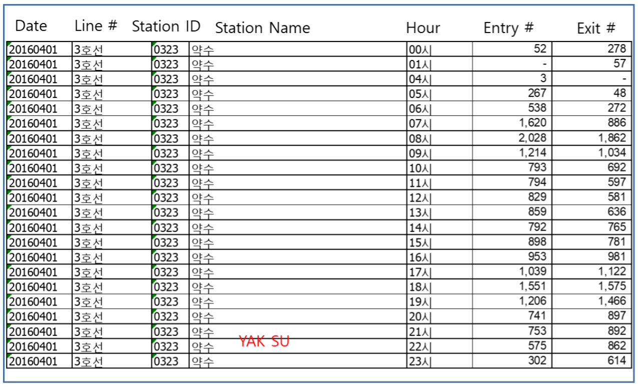
We have demonstrated our underlying idea for solving the inverse problem by applying the mass conservation law using a simple example. Owing to the simplicity of the example, we did not introduce our optimization solver. Our full scale model with optimization is described below.

As presented in Appendix, the subway map of Seoul city shows that the subway system has nine lines with 291 stations. The sizes of the station datasets are summarized in Table 5. The total number of records in the database is approximately $291 \times 365 \times 24 = 2,549,160$.

TABLE 5. Data summary of Seoul metro subway.

Parameter	Value
Number of lines	9
Number of stations	291
Period of dataset	Mar. 2016 - Feb. 2017
Number of records per day per station	$<$ 24 (hourly)
Total count of entry passengers	$1,876,043,806$
Total count of exit passengers	$1,873,686,411$

Figure 1 presents a snapshot of the raw data obtained from the Department of Transportation (DOT) of Seoul City, showing that the type of database is relational and that the records consist of seven columns of parameters. The first row can be interpreted as follows: At hour 0 on April 1, 2016, 52 passengers entered and 276 exited the gates at the station. For demonstration purposes, we selected the *YAKSU* station on Line 3, which is located in a residential area near the *Han River* in Seoul City. *YAKSU* is a station located in a typical residential area (courtesy of D.O.T of Seoul city, R.O.Korea).



Date	Line #	Station ID	Station Name	Hour	Entry #	Exit #
20160401	3호선	0323	약수	00시	52	278
20160401	3호선	0323	약수	01시	-	57
20160401	3호선	0323	약수	04시	3	-
20160401	3호선	0323	약수	05시	267	48
20160401	3호선	0323	약수	06시	538	272
20160401	3호선	0323	약수	07시	1,620	886
20160401	3호선	0323	약수	08시	2,028	1,862
20160401	3호선	0323	약수	09시	1,214	1,034
20160401	3호선	0323	약수	10시	793	692
20160401	3호선	0323	약수	11시	794	597
20160401	3호선	0323	약수	12시	829	581
20160401	3호선	0323	약수	13시	859	636
20160401	3호선	0323	약수	14시	792	765
20160401	3호선	0323	약수	15시	898	781
20160401	3호선	0323	약수	16시	953	981
20160401	3호선	0323	약수	17시	1,039	1,122
20160401	3호선	0323	약수	18시	1,551	1,575
20160401	3호선	0323	약수	19시	1,206	1,466
20160401	3호선	0323	약수	20시	741	897
20160401	3호선	0323	약수	21시	753	892
20160401	3호선	0323	약수	22시	575	862
20160401	3호선	0323	약수	23시	302	614

FIGURE 1. Snapshot of raw data (hourly logs) on entering and exiting passengers at *YAKSU* station on April 1, 2016. The total entries and exits are 17, 807 and 16, 898, respectively.

We adopt the conservation law of mass based on a fairly strong assumption that a passenger will return to a particular station. As seen in Figure 1, the numbers of entries and exits at the *YAKSU* station for April 1, 2016 were 17, 807 and 16, 898, respectively. The two numbers are weakly equal at the 95% confidence level.

Figure 2 provides the frequency graphical representation of entry and exit data recorded in Table 5. For convenience of presentation, the blue and red curves indicate entry and exit data, respectively. Both of the colored curves show peaks in the morning and evening. The graph also reveals that the blue curve indicates a higher frequency level from 06:00 to 16:00, whereas the red curve indicates that the higher frequency is from 17:00 until 02:00. This seems to indicate that passengers leave in the early morning from home and return after spending a certain number of hours outside of the subway.

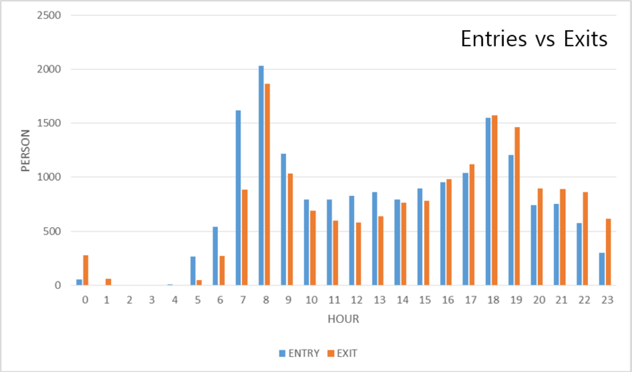


FIGURE 2. Frequency representation of entry and exit data. The *X* axis indicates hours (from 00:00 to 23:00) and the *Y* axis is the number of passengers.

Each of the colored frequencies can be decomposed into a series of unitary functions. We propose a unitary function that is a combination of two delta functions [58].

$$U(\tau, \triangle T) := \delta_t(\tau; i) + \delta_{t+\triangle T}(\tau; e) \tag{10}$$

Here, δ is the Dirac-Delta function and the second arguments i and e indicate entries and exits, respectively. It is reasonable to state that the frequency data are observational data from the conservation law of mass. This implies that the frequency data can be decomposed using the unitary function of Eq.10. The unitary $U(t)$ value is designed to fit the data shown in Table 5, where the table consists of two columns: 'entry' and 'exit'. Using $U(\tau, \triangle T)$, $\delta_T(\tau; i)$, we can fit the blue graph (entry data) and $\delta_{T+\triangle T}(\tau; e)$ is used to fit the red graph (exit data) from Figure 2.

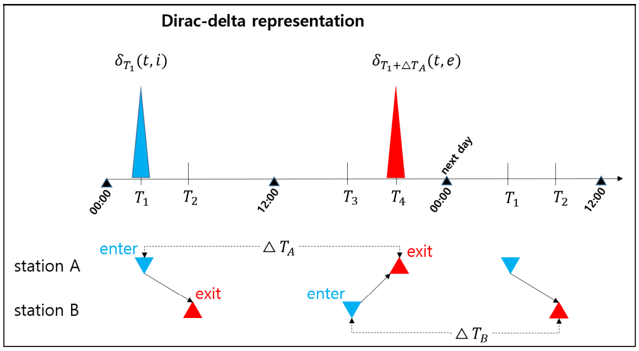


FIGURE 3. Illustration of Eq.10 shows definition of a unitary function *U* as the combination of two Dirac-delta functions, δ_τ and $\delta_{\tau+\Delta T}$; the first term is for entry ridership and the second is for exit ridership, respectively.

Figure 3 explains the Dirac-Delta representation of Eq. 10 by illustrating the temporal procedure of a round-trip passenger: a passenger enters subway station A at hour T_1, exits subway station B at T_2, re-enters station B at T_3 after some time, and returns to station A at T_4. As a result, the passenger spends $\triangle T_A (= T_4 - T_1)$ hours before returning to station A. The blue peak signal bar at the left and the red one at the right illustrate the two Dirac-delta functions defined in Eq.10. At T_1, the entry ridership count of station A is increased by 1, and at T_4, the exit ridership count of station A is increased by 1. On the other hand, from the perspective of station B, this round trip can be interpreted another way: a passenger enters subway station B at time T_3 and returns at time T_2 the next morning after spending $\triangle T_B (= 24 - T_3 + T_2)$ hours elsewhere.

The summation of $U(\tau, \triangle T)$ for all the passengers will match the columns for entry and exit data. Consider the following formula.

$$\Psi(s, d) = \sum_{d \in D} \sum_{s \in S} \sum_{\triangle T=0}^{23} \sum_{\tau=0}^{23} \boldsymbol{a}_{\tau,\triangle T} \cdot \boldsymbol{U}(\boldsymbol{\tau}, \triangle \boldsymbol{T}) \qquad (11)$$

Here, D and S denote the date and station, respectively, and $a_{\tau,\triangle T}$ is constant. It is straightforward that $\Psi(s, d)$ matches the ridership table data by its definition if (s, d) indicates a single station and single date, respectively. However, if we expand the range of (s, d) to all stations and all available dates, then $\Psi(s, d)$ is an overdetermined system consisting of several linear equations with 24×24 unknown parameters$(a_{\tau,\triangle T})$. Eq.11 is a typical optimization problem. We introduce the following proposition to prove the solvability of our formula.

Proposition 1: For any given period of dates and a certain set of stations, it is feasible to calculate $a_{\tau,\triangle T}$ for $\Psi(s, d)$, which satisfies the following constraint equation:

$$\operatorname{argmin}\left(a_{\tau,\triangle T}\right) \{|\Psi - \mathbf{R}|\}, \qquad (12)$$

where Ψ is defined in Eq.11, $\mathbf{R}$ is ridership data, and τ, $\triangle T = 0, 1, \ldots, 23$.

Proof: Daily ridership at a given station is expressed by two discrete functions as follows:

$$f = \{f[0], f[1], \ldots, f[23]\}, \quad \text{for entry data}$$
$$g = \{g[0], g[1], \ldots, g[23]\}, \quad \text{for exit data}$$

where f and g represent the blue and red curves shown in Figure 2, respectively. Daily ridership for a specified group of stations (S) and for a given period of dates is expressed as follows:

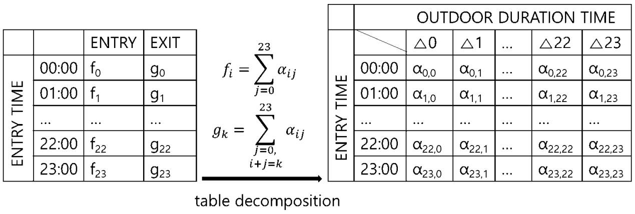
$$F(S, D) = \{f \,|\, \text{station} \in S \text{ and date } \in D\}$$
$$G(S, D) = \{g| \text{ station } \in S \text{ and date } \in D\}$$

We denote $\mathbf{R}(S, D) := \{F(S, D), G(S, D)\}$, where $\mathbf{R}$ corresponds to the Seoul Metro subway ridership data during March, 2016 to Feb. 2017. We constructed a computational optimization model as described by Eq.12 and adopted a heuristic GA algorithm for a computation simulation of Eq.12, which is described in detail in the next section. First, we explain the property of the conservation law of mass by integrating it with the wave decomposition method as follows.

Here, we present a wave decomposition process using the elementary function $U(\tau, \triangle T)$ (see Figure 3 and Eq.12) from raw ridership log data. One round trip of a passenger, from A to B and back from B to A, adds a ridership of up to four in the following manner: add 1 to the entry log at A, exit log at B, entry log at B, and exit log at A. The entry log time at A was T_1 and the exit log time at A after the passenger returned was $T_4 = T_1 + \triangle T_A$. The ridership changes involved in this trip episode can be represented by the elementary function $U(\tau, \triangle T)$ defined in Eq.11. In this manner, the two columns of entry and exit data can be written as 24 × 24 summations of

$$\sum_{\tau=0}^{23} \sum_{\triangle T=0}^{23} a_{\tau,\triangle T} \cdot U(\tau, \triangle T).$$

The $a_{\tau,\triangle T}$ values denote 24×24 matrix elements, which are illustrated in Figure 4.



		ENTRY	EXIT
ENTRY TIME	00:00	f_0	g_0
	01:00	f_1	g_1
	...	...	...
	22:00	f_{22}	g_{22}
	23:00	f_{23}	g_{23}

$$f_i = \sum_{j=0}^{23} \alpha_{ij}$$

$$g_k = \sum_{\substack{j=0, \\ i+j=k}}^{23} \alpha_{ij}$$

	OUTDOOR DURATION TIME					
		△0	△1	...	△22	△23
ENTRY TIME	00:00	$\alpha_{0,0}$	$\alpha_{0,1}$	...	$\alpha_{0,22}$	$\alpha_{0,23}$
	01:00	$\alpha_{1,0}$	$\alpha_{1,1}$	...	$\alpha_{1,22}$	$\alpha_{1,23}$
	...	...	...	...	...	...
	22:00	$\alpha_{22,0}$	$\alpha_{22,1}$	...	$\alpha_{22,22}$	$\alpha_{22,23}$
	23:00	$\alpha_{23,0}$	$\alpha_{23,1}$	...	$\alpha_{23,22}$	$\alpha_{23,23}$

FIGURE 4. Table decomposition. The entry and exit tables are decomposed by entry time based on outdoor duration length. In the right-side table, α_{ij} represents the number of passengers who entered at the i-th hour and exited at the $(i + j)$-th hour through a gate of station A.

A Further description of decomposition follows. As seen in Figure 4, a number f_0 of passengers entered station A between 00:00 and 01:00, and returned to exit A after spending a few hours outside the station. Their exit records must be added to one of the $\{g_0, \ldots, g_{23}\}$. Among the passengers, a certain number, $\alpha_{0,0}$, came back within 1 hour (by 01:00), $\alpha_{0,1}$ within 2 hours (by 02:00), and later than 02:00, $\cdots$, and finally a certain number, $\alpha_{0,23}$, returned within 24 hours. It is natural to think that the number f_0 was decomposed into α's, where $f_0 = \sum_{j=0}^{23} \alpha_{0\,j}$. Similarly, $f_i = \sum_{j=0}^{23} \alpha_{ij}$ for each hour $i = 0, \cdots, 23$. The decomposed table can be considered as a 24×24 matrix. In this manner, daily ridership of a station is transformed into a matrix. Note that the table at the right represents ridership for a specific station and date. In general, the values of the elements, α_{ij}, are variables representing the station and date, that is, $\alpha_{ij}(d, s)$.

Eq.13 seems to apply an unrealistically strong assumption for construction of the conservation model. We validate that this assumption is acceptable at the end of §VI based on results using ground truth data.

$$F := \sum_{i=0}^{23} f_i = \sum_{i=0}^{23} g_i =: G \qquad (13)$$

Under the assumption of $F = G$, the matrix representation $[\alpha_{ij}]$ for the wave decomposition of ridership is straightforward. Furthermore we introduce a normalized matrix Ω from the decomposed table $[\alpha_{ij}]$.

$$\Omega = \begin{bmatrix} \omega_{0,0} & \cdots & \omega_{0,23} \\ \vdots & \ddots & \vdots \\ \omega_{23,0} & \cdots & \omega_{23,23} \end{bmatrix} \tag{14}$$

Here,

$$\omega_{i,j} = \frac{\alpha_{i,j}}{\sum_n \sum_m \alpha_{m,n}}.$$

Therefore the coefficient value $\alpha_{i,j}$ from Ω corresponds to Eq.15 below.

$$\alpha_{ij} = F \cdot \omega_{ij} \tag{15}$$

We claim that the matrix Ω can thus represent the characteristics of daily ridership. □

V. COMPUTATIONS

The coefficient of unitary function $U(\tau, \triangle T)$, appearing as $a_{\tau,\triangle T}$ in Eq.11 actually corresponds to α_{ij} in Eq.15, where $\tau = i$ and $\triangle T = j$. Therefore, solving Eq.12 is equivalent to finding Ω. To compute the matrix Ω, our approach is to employ an approximation solver for the probabilistic optimization formula of Eq.12. The degree of freedom of the problem is non-trivial; it is $24 \times 24 (= 576)$. The high degree of freedom of the parameter space led us to choose an iterative method using a heuristic GA-based method.

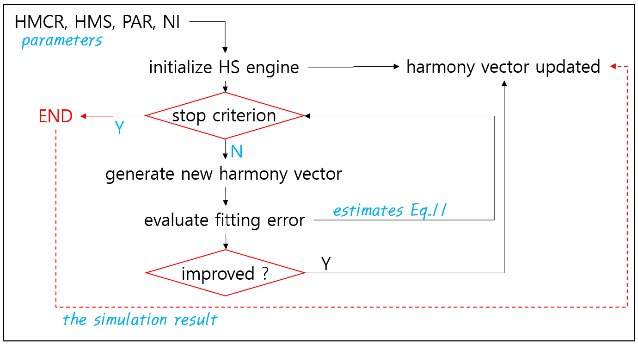


FIGURE 5. Harmony search algorithm adopted in this study, which consists of four components: 1) engine initialization, 2) harmony vector generation, 3) evaluation of fitting error, and 4) update solution. Among the four components, the third (i.e., evaluation of the fitting error) is the specific problem, whereas the others are general operations.

Among various heuristic GA-based methods, we adopted the harmony search algorithm designed to solve global optimization problems [48]–[52]. Figure 5 provides an overview of the algorithm applied in this study. The algorithm consists of four parts: initialization of the engine, generation of the harmony vector, evaluation of the fitting error, and generating an updated harmony vector. At the step of generating a (new) harmony vector, we first created a 24×24 dimensional random vector with a range of [0, 1], and then normalized the vector to fit scales with the matrix Ω. Of the four components, evaluation of the fitting error was the specific problem, whereas the others were general operations. In this study, the third component, namely, the evaluation of the fitting error, is computed by the formula defined in Eq.12.

TABLE 6. Summary of terms and parameter settings for harmony search engine.

NVAR	Number of variables	24×24
HMS	Harmony memory size	2, 400
MaxIter	Maximum iterations	100, 000, 000
PAR	Pitch adjusting rate	0.4
BW	Bandwidth	0.2
HMCR	Harmony memory consideration rate	0.9

As previously mentioned, the third component (error fitting) is the only problem-specific operation. A pseudo-code for the fitting function is provided in Table 6. The fitting error evaluation reads the raw subway ridership data to assess the total daily passengers and applies Eq.14 to build matrix Ω. In this manner, the discrepancy between the real raw data and synthetic data generated from Ω is estimated by the least squares measure.

Once the fitting error is estimated, the harmony search engine compares the history of fitting errors to determine whether the newly generated harmony vector is improved. If improved, the harmony vector is updated by the new vector.

We focused on describing the harmony search algorithm as an optimization solver used in this study to generate the best fitting Ω_{ij}. Several typical parameters required for the harmony search engine are summarized in Table 6.

As seen in the table, NVAR was set to 24×24 because we take elements of Ω as stochastic variables to fit. Harmony memory size (HMS) was set to 2, 400, a rather large number, because we must consider a large degree of freedom (24×24). The number of maximum iterations for the stopping criterion was also set as a large number, namely, 100, 000, 000. The internal settings were typically as follows: pitch adjusting rate (PAR) = 0.4, bandwidth (BW) = 0.2, and harmony memory consideration rate (HMCR) = 0.9. Figure 6 presents a graph of the fitting errors calculated by the optimization solver adopted in this study at each iteration step. In real simulations, MaxIter is set as 100, 000, 000, a stopping criterion of the iterative solver, but for the figure we limited the range to 300, 000 for visualization purposes.

In Figure 6, the blue curve indicates the fitting error at each iteration step calculated by the GA-based heuristic search algorithm, and the red curve is the best fitting error produced by the harmony search algorithm. The error value indicates discrepancies between the observed and simulated ridership. The fitting error is evaluated as

$$\epsilon := \frac{\|\Psi(s,d) - R^s(d)\|}{n(S) \cdot n(\text{days})}, \tag{16}$$

Algorithm 1 Pseudo-Code for Fitness Function of Harmony Search Algorithm: Fitness(xNrm)

```
Require: xNrm: 24x24 normalized vector generated by HS random
  generator.
Require: mapDate2Station2Data: a dictionary object constructed
  from the raw data in form of ridership per station per day. date →
  station → data.
  # allocation
  #   build probability tables
  #   data type: array~\hbox{[24]}
  xpdfInHourly ← array of 0, indexed from 0 to 23
  xpdfOutHourly ← array of 0, indexed from 0 to 23

  # initialization
  row ← 0
  for row < 24 do
    # 'row' indicates the entry hour
    # 'countInPdf' denotes # of passengers at the hour
    countInPdf ← 0
    col←0
    for col < 24 do
      # 'col' indicates the exit hour
      inc ← xNrm[row * 24 + col]
      countInPdf += inc
      xpdfOutHourly[(col+row)%24] += inc
      col ← col + 1
    end for
    row ← row + 1
  end for
  # initialize 'fit'
  fit ← 0
  # loop a Map(key, value)
  #   mapDate2Station2Data,
  #     where key -> date and value -> map<station, data>
  #
  # for each day
  for all date in mapDate2Station2Data.key do
    # assign 'station to data' for the date selected
    mapStation2Data ← mapDate2Station2Data[date]
    # iterate through stations
    for all station in mapStation2Data.key do
      # obtain the ridership data at the station
      dataList ← entrySt.getValue();
      # last element of ridership is the total counts
      cardinal ← dataList.fileDataInp[23];
      for k < 24 do
        # hourly IN passengers
        dif ← dataList.fileDataInFreq[k] - xPdfInHourly[k] *
        cardinal;
        fit ← fit + (dif * dif);
        k ← k + 1
      end for
      k ← 0
      for k < 24 do
        # hourly OUT passengers
        dif ← dataList.fileDataOutFreq[k] - xPdfOutHourly[k] *
        cardinal;
        fit ← fit + (dif * dif);
        k ← k + 1
      end for
    end for
  end for
  # return mean squared error
  return $\sqrt{fit}$
```

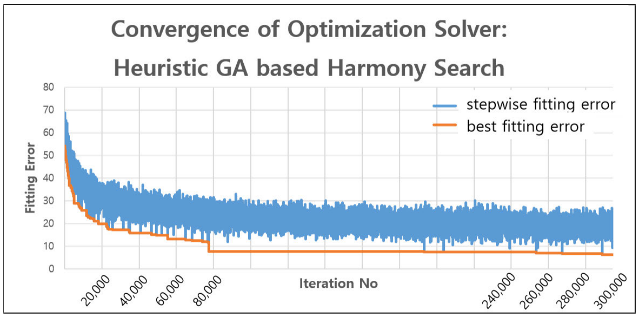


FIGURE 6. Convergence curve of iterative optimization solver adopted in this study. Table 6 describes the selection of parameters. The *X*-axis indicates the iteration number and the *Y*-axis indicates the fitting error value denoted by ϵ, as defined in Eq. 16.

where Ψ denotes the simulated value as defined in Eq.12, T is the observation data as defined in Eq. 5, S indicates the set of subway stations as also defined in 5, and ϵ is the average discrepancy at a station per day between the observation data and the simulation result. The red graph shows that stability is achieved at approximately $80,000$ iteration steps.

VI. RESULTS

In this section, we present the simulation results of our computational model using Eq.12 applied to a real dataset, the Seoul metro subway ridership database for the period of Mar. 2016 to Feb. 2017.

Considering Ω as an output matrix (of size 24×24) obtained from the wave decomposition of daily ridership data at a specific station and date, the total of all R's generated from raw data is 291×365 ($= 106,215$) for all stations and dates. Then, the harmony search algorithm is adopted to generate and find the best fitting Ω to satisfy the optimization problem formulated in Eq.12. The matrix Ω, which is the main output of our model, is obtained over a very long period, specifically, $108,865.6$ seconds (approximately 30 hours), or an average of $100,000,000$ iterations, before a stopping criterion is satisfied.

A result of the simulation is provided in Figure 7. The rows and columns of the matrix in the figure indicate the return time $\triangle T$ and entry hour (τ), respectively. The matrix Ω is presented in the form of a heat map for visualization purposes. As seen in Eq.14, the range of values for $w_{ij} \in \Omega$ is $[0, 1]$. The color map varies from 0 ($[0, 255, 0]$; green color) to 1 ($[255, 0, 0]$; red color), where red cells imply high probability.

From Figure 7, we can observe that the reddish regions are clustered into two areas. The left area explains a passenger's original routine: $A_{\text{in}} \rightarrow B_{\text{ex}} \rightarrow \text{out} \rightarrow B_{\text{in}} \rightarrow A_{\text{ex}}$, where A and B denote subway stations. The right part reflects a passenger's mirror routine: $B_{\text{in}} \rightarrow A_{\text{ex}} \rightarrow \text{home} \rightarrow A_{\text{in}} \rightarrow B_{\text{ex}}$, from the point of view of station B. Henceforth, if we consider $\triangle T$ in the left part as the **ODT** of a passenger, then we can interpret the right area as the home duration time (**HDT**) of a passenger. For simplicity of presentation, however, we omit the similar comprehensive analysis of **HDT**.

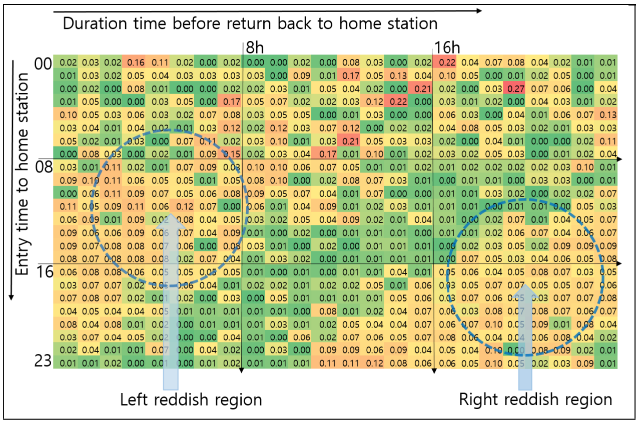


FIGURE 7. Heat map representation of decomposition matrix Ω of a station specified as the 'home station' (in this study, it is the *YAKSU* station. Columns indicate entry time to the home station and rows specify the elapsed hours before returning to the home station. The value of each matrix element indicates the probability of returning to the home station after *j* hours. The heat map values are clustered into the two regions denoted by red circles.

From Figure 7, we can further observe the correlative property between the consecutive row vectors of matrix Ω. The $j+1$-th row looks similar to the j-th row after moving it one step in the left direction, i.e., $\Omega(j, i) = \Omega(j+1, i-1)$ for $i = 1, 2, .., 23$. Therefore, Figure 8 is constructed to check the correlation. In the left part of the figure, the X axis indicates the row order (i.e., j means the j-th row) and the Y axis represents the correlation coefficient value. We are interested in the step-size, where CORR△0, CORR△1, CORR△2 indicate a 0-step, 1-step, and 2-step move to the left, respectively. The blue curve illustrates the correlation between two consecutive rows without any move (i.e., CORR△0), the amber curve shows the correlation coefficient between two consecutive rows for a forward transition to the left by one step of the latter row (i.e., CORR△1), and the gray curve shows the value for a two-step forward transition of the latter row (i.e., CORR△2). In the right part of the figure, the change represents the movement of column vectors of Ω instead of row vectors. The transition to the left used for the row vectors is replaced with an upward movement for the column vectors.

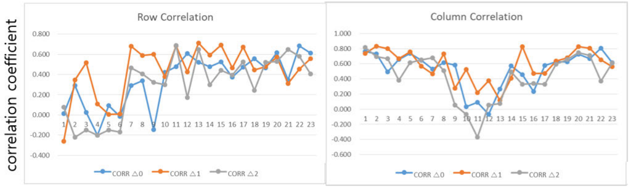


FIGURE 8. The left side shows row-vector correlations and the right shows column-vector correlations of matrix Ω. The *X* axis indicates the order of rows and columns for the top and the bottom, respectively. The *Y* axis indicates the correlation coefficient. Descriptions of the CORR△0, CORR△1, and CORR△2 notations are explained in the text body.

As seen in Figure 8, the amber colored graphs show the best correlativity for both row and column vectors. This can be interpreted as

$$\omega_{i,j} = \omega_{i+1,j-1}.$$

In other words, the probability representing a passenger who departed at 08:00 and returned at 18:00 (i.e., $\omega_{8,10}$) equals to $\omega_{9,9}$, which is the probability of a passenger departing at 09:00 and returning at 18:00) after 9 hours duration time. This implies that the passenger **ODT** has strong dependency on entry time. In short, an earlier departure implies a longer duration time.

We have discussed the dependency of the entry hour to the subway station on the **ODT**. We next consider the dependency of days of the week. As clearly shown in Figure 9, which compares Sunday and Monday ridership, the weekday ridership shows the three clustered red regions, whereas the weekend shows only two. In contrast to the weekend heat map, in the Monday heat map, low ridership is apparent between the 10th and 15th hours. It is reasonable to state that subway passengers go to work in the morning hours before the 9th hour, thus indicating the difference between Sunday and Monday ridership.

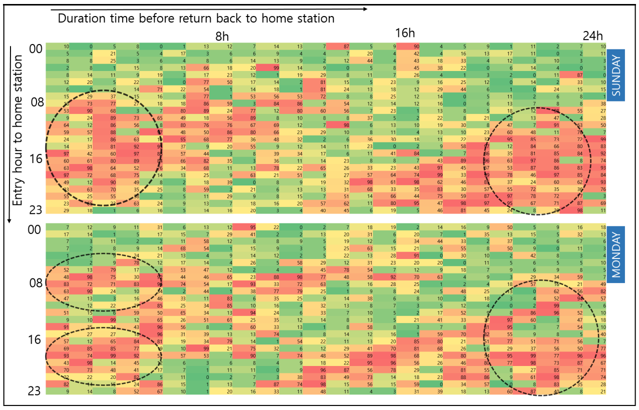


FIGURE 9. Heat maps of entry time vs. ODT (outdoor duration time) of passengers. The top and the bottom refer to Ω for Sunday and Monday, respectively. The shapes of the red clustered regions are different.

We take the example of Monday for weekdays in the previous paragraph. For the other days, Figure 17 provides a weekday heat map representation of Ω and specifically reveals the three red clustered regions that weekdays have in common.

Although it is useful to show discrepancies visually, the heat-map itself lacks statistical details. We present a column-wise projection profile of the heat maps in Figure 10. The figure illustrates the average times passengers remain away from home before returning to the station from which they departed. The graphs reveal the differences between the maximum frequency between Sunday and Monday ridership; a 2 hour **ODT** is Sunday's maximum, whereas a 12 hour **ODT** is Monday's. The dotted line is the 1-step moving average, which reveals that passengers stay away from home for a shorter period on Sunday than on Monday.

We have proposed our model in terms of the conservation law and stochastic optimization theory, and have presented the results based on the big ridership data of the Seoul Metro subway system. As mentioned briefly, the model requires an

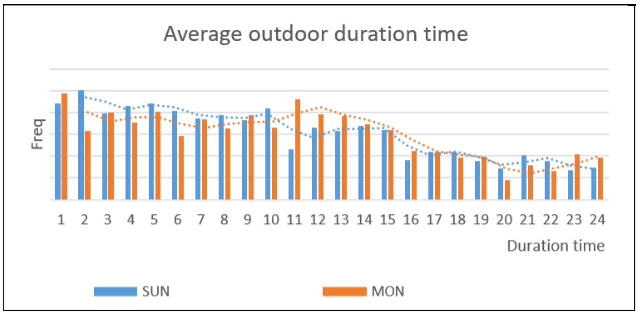
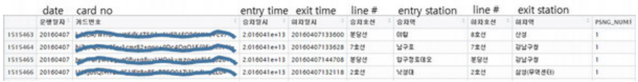


FIGURE 10. Projection profile of Ω for comparison of average **ODT** (outdoor duration time) graphs obtained by the column sum (average) of Ω in Figure 9. Sunday and Monday graphs are blue and red, respectively. *X* axis indicates **ODT** in hours. *Y* axis is the column average of Ω. The dotted line is the 1-step moving average, which shows that the **ODT** of Sunday is shorter than the **ODT** of Monday within the time region [0, 12].



	date 운행일자	card no 카드번호	entry time 승차일시	exit time 하차일시	line # 승차호선	entry station 승차역	line # 하차호선	exit station 하차역	PSNG_NUM
1515463	20160407	[redacted]	2.016041e+13	20160407133600	분당선	야탑	8호선	산성	1
1515464	20160407	[redacted]	2.016041e+13	20160407133628	7호선	남구로	7호선	강남구청	1
1515465	20160407	[redacted]	2.016041e+13	20160407144708	분당선	압구정로데오	분당선	강남구청	1
1515466	20160407	[redacted]	2.016041e+13	20160407132118	2호선	낙성대	2호선	삼성(무역센터)	1

FIGURE 11. Snap shot of transit record list with traffic card IDs (covered over) for Seoul Metro subway system.

assumption that all passengers take round trips and that they return to the station from which they departed earlier, which is a strong assumption. Aside from the ridership data, Seoul Metro subway system also manages transit records containing traffic card information. A snapshot of the database is presented in Figure 10. This data is strictly prohibited to the public because it contains private information including card number. We were only allowed to obtain 1 week's data (from April 1st to April 7th in 2016) which amounted to 7 GB in CSV file format, and had to return the data to the authority after a two-week research period.

TABLE 7. Descriptive statistical analysis for the data illustrated in Figure 11. Single rider indicates a passenger whose card appears only once in the list of transit records for a day. Round trip means a passenger whose card appears twice. Multiple rider denotes a passenger whose card appears more than three times.

Date		Total	Single	Round Trip	Multiple
April. 1	FRI	$8,802,017$	17.80%	57.34%	24.86%
April. 2	SAT	$6,919,048$	20.76%	53.92%	25.32%
April. 3	SUN	$4,641,016$	24.80%	54.10%	21.10%
April. 4	MON	$8,414,771$	17.00%	61.62%	21.38%
April. 5	TUE	$8,528,197$	16.06%	60.72%	23.22%
April. 6	WED	$8,612,054$	16.58%	60.08%	23.34%
April. 7	THU	$8,447,590$	17.44%	60.46%	22.10%
AVG		$7,766,384$	18.63%	58.32%	23.04%

The data consisted of $54,364,693$ transit records, where each record indicates a single ride (entry and exit) by a passenger. For each day, we applied matching of traffic card IDs and present the output in Table 7. The results show that only 58% of passengers take round trips, and 18% are single riders who are not regular users of Seoul Metro. Approximately 23% of users are multiple riders, who take more than three trips per day. We could not identify what is these population.

The objective of this analysis is to identify the rate percentage of passengers who return to the station from which they departed, namely *PRR* (passengers' return rates). Table 8 summarizes the results. The overall *PRR* are 54.28% on average, and of this percentage, round trip passengers represent 93.04% on average.

TABLE 8. Return rates of passengers (*PRR*) using Seoul Metro subway. The 'Overall' column indicates the average return rate, and the 'Round Trip' column indicates the return rate among passengers taking round trips.

Date		Overall	Round Trip
April. 1	FRI	52.68%	91.87%
April. 2	SAT	48.78%	90.47%
April. 3	SUN	50.88%	94.05%
April. 4	MON	58.14%	94.35%
April. 5	TUE	57.36%	94.30%
April. 6	WED	56.04%	93.28%
April. 7	THU	56.08%	92.96%

Our model would not be a good fit for overall passengers owing to a lack of the mass conservation property (almost half of the passengers do not return to the same station). However if we set the constraint conditions on the round trip passengers, then our model is well defined.

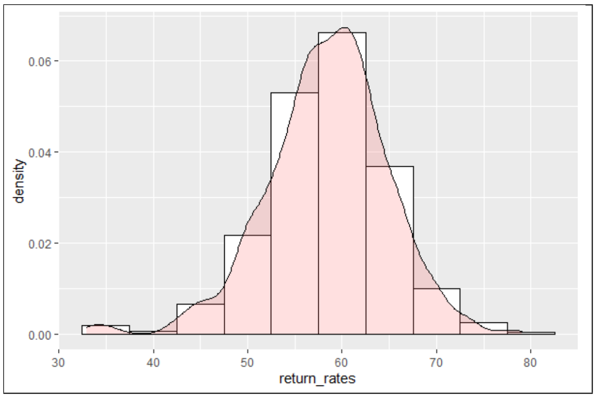


FIGURE 12. Distribution of passengers' return rates at the stations. *X*-axis denotes *PRR* and *Y*-axis indicates frequency density. The size of bins is 5.

Figure 12 presents the frequency histogram of *PRR* (with bin size of 5) at the stations. The range of *PRR* is 33% to 79%. There are 12 stations having *PRR* less than 45% which can be characterized as follows: 2 stations of them are airports (Gimpo and Inchon); 5 of them are popular outdoor camping places; 3 of them are is famous street market places especially for the foreign traveller; and 2 of them are university stations located at suburban.

A. VALIDATION

Our model Ψ extracts **ODT** at each hour from hourly accumulated ridership data without transit record details. Mathematical stability of Ψ basing on the optimization model described in proposition 1 is guaranteed by the property of conservation law of mass (Eq.6). System stability on the input parameters (R^s_d) is also guaranteed by a heuristic GA based HS algorithm as shown in Figure 6.

On the other hand, validation of the experimental results is achieved by matching with the ground-truth data. Our main experimental result **ODT** representing the matrix Ω is the target of validation. Denote Ω^Ψ and Ω^G as **ODT** matrices by our model simulation and the ground-truth, respectively. $L^1, L^2, \text{and} L^\infty$ of $|\Omega^\Psi - \Omega^G|$ are not good measures for comparison of the two normalized matrices due to high degree of freedom (24 × 24 number of elements). The higher degree of freedom tends to the higher R^2. Instead, average **ODT** (average sum of 24 counts of hourly **ODT**) is used for validation. The average **ODT**, a 24 dimensional vector, is a column sum of Ω as seen at the left panel in Figure 8.

We explain the details of the validation process with an example case of the *YAKSU* station, and present the statistics of all stations. Figure 13 illustrates comparison of the two averaged **ODT**s from the simulation (Ψ) and from the ground-truth (R).

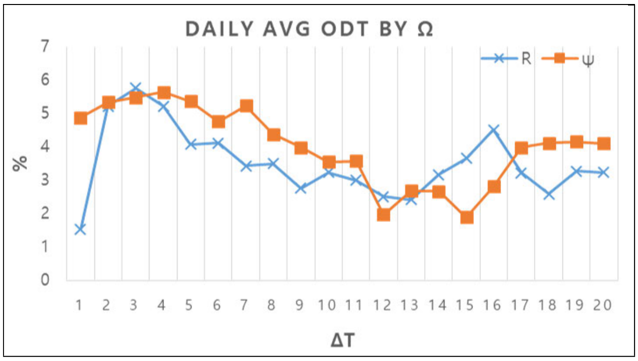


FIGURE 13. Illustration of daily ODT comparison between simulation (amber) and ground-truth (blue). *X*-axis indicates ODT in elapsed hour and *Y*-axis denotes frequency in percentage. The typical 'R-squared' value is used as a measure of discrepancy.

Figure 14 illustrates scatter plots of the points of curves in Figure 13. For a point $P = (p_x, p_y)$ in Figure 14, p_x and p_y are **ODT** of ground-truth and simulated, respectively. The four different tones of gray colors indicate four different data sets compared. The lightest color plots are from the same data used in Figure 14. Inclination of the linear trend line and formation of the scattered points show correlativity between the ground-truth and the simulated.

For the discrepancy analysis, we applied the linear regression test to obtain ’R-squared’ value. An exemplary summary output is presented below, where, in the call ’lm(formula = odt_g_n odt_s_n)’, odt_g_n and odt_s_n denote the average **ODT** obtained from the ground truth and from the simulation, respectively. In the summary output, residual standard error (RSE) is 0.6887, ’R-squared’ is 0.5829, p-value is 0.0002273, and F-statistic is 22.26 with 16 DF (degree of freedom). Those numbers indicate that the responses depend strongly on the observations.

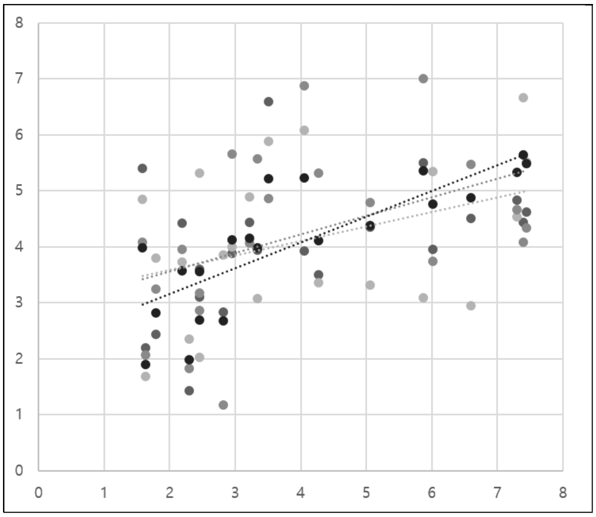
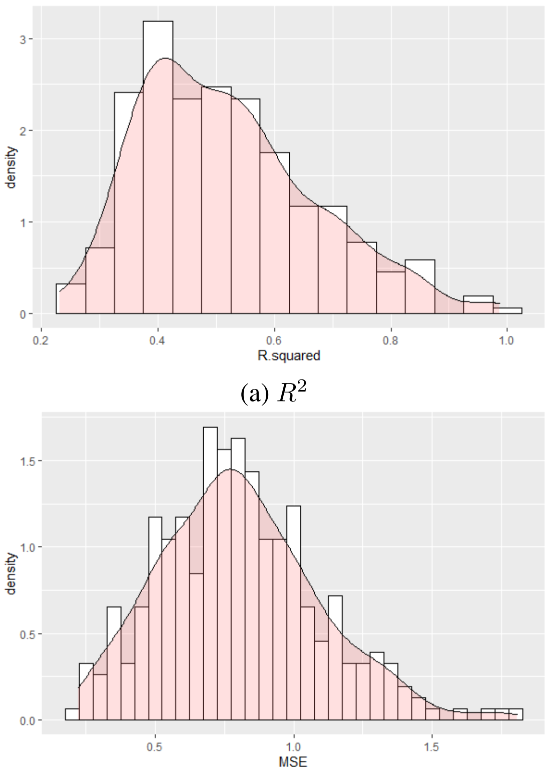


FIGURE 14. Illustration of daily ODT comparison between simulation and ground-truth. For a point $P = (p_x, p_y)$ in the graph, p_x indicates ground truth value and p_y denotes simulated value from the curve in Figure 13.

```
                      output of ’lm’ function
Call:
lm(formula = odt_g_n ~ odt_s_n)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9165 -0.4731 -0.1177  0.4787  1.1235

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10695    0.36606  19.414 1.51e-12 ***
odt_g_n     -0.27925    0.05906  -4.728 0.000227 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05
                ’.’ 0.1 ’ ’ 1

Residual standard error: 0.6887 on 20 degrees of
                                         freedom
Multiple R-squared:  0.5829,
Adjusted R-squared:  0.5568
F-statistic: 22.36 on 1 and 20 DF,
p-value: 0.0002273
```

From the summary output of ’lm’, we selected ’Residual standard error’ (*RSE*) and ’Adjusted R-squared’ (R^2) as validation measures for all the 291 stations. *RSE* measures the average amount that the response deviates from the ground-truth. R^2 is proportion of variance: 0 means a regression does not explain the variance of the response variable, and 1 explains the variance (e.g., in the above example, 55% of the variance found in ground-truth (odt_g_n) can be explained by the variable of simulated outputs (odt_s_n). Figure 15 presents the probability distribution of R^2 and *RSE* over the stations (up).

The snippet (’QQ ranges of RSE and R-squared’) below is summary of the simple statistics of *RSE* and R^2 of 291



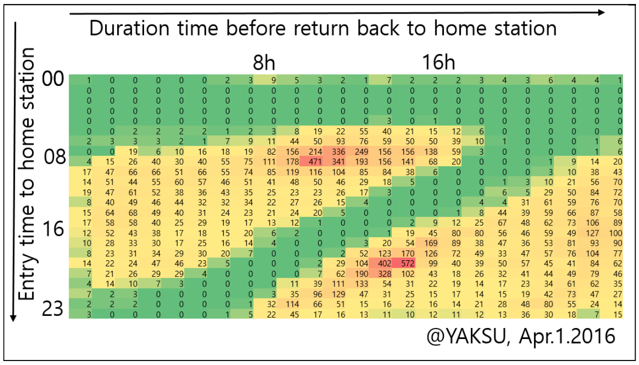
(a) R^2

(b) *RSE*

FIGURE 15. Distribution of R^2 and *RSE* over 291 stations. *X*-axis indicates R^2 (a) and *RSE* (b) values and *Y*-axis denotes normalized density. The sizes of bin are .05 for the both graphs.

stations. There are no strict criteria on the threshold values of *RSE* and R^2. From the point of view of *RSE*, the amount 0.9789 (the third quartile value) for the mean square sum of residual errors is small enough to be a threshold. For *RSE*, we normalized the two **ODT**s and used 22 data points (excluding the first and the $24-th$ hour duration time). On the other hand, from the point of view from R^2, the amount 0.4049 (the first quartile value) is high enough to be a threshold. 40% of variance in response variable can be explained by the variance in ground-truth data. Consequently, the simulation outputs of our model, **ODT**, are validated for at least 75% of the stations. Our model did not produce meaningful outputs from some of the stations having non-regular passenger's transit records.

```
            QQ ranges of RSE and R-squared
       RSE                R.squared
Min.   :0.2238    Min.   :0.2318
1st Qu.:0.5921    1st Qu.:0.4049
Median :0.7858    Median :0.5007
Mean   :0.8031    Mean   :0.5180
3rd Qu.:0.9789    3rd Qu.:0.6025
Max.   :1.8078    Max.   :0.9870
```



									8h								16h							
00	1	0	0	0	0	0	2	3	9	5	3	2	1	7	2	2	2	3	4	3	6	4	4	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0
	0	0	2	2	2	2	1	2	3	8	19	22	55	40	21	15	12	6	0	0	0	0	0	0
	2	3	3	3	2	1	7	9	11	44	50	93	76	59	50	50	39	10	1	0	0	0	1	6
	0	0	19	6	10	16	18	19	82	156	214	336	249	156	156	138	59	3	0	0	0	0	1	6
08	4	15	26	40	30	40	55	75	111	178	471	341	193	156	141	68	20	0	0	0	1	9	14	20
	17	47	66	66	51	66	55	74	85	119	116	104	85	84	38	6	0	0	0	0	3	10	38	43
	14	51	44	55	60	57	46	51	41	48	50	46	29	18	5	0	0	0	1	3	10	21	56	70
	19	47	61	52	38	36	43	35	25	23	23	26	17	3	0	0	0	0	6	5	29	50	84	72
	8	40	49	46	44	32	32	34	22	27	26	15	4	0	0	0	0	4	4	31	61	59	76	70
	15	64	68	49	40	31	24	23	21	24	20	5	0	0	0	0	1	8	44	39	59	66	87	58
	17	58	58	40	25	29	19	17	13	12	1	0	0	0	2	9	12	25	67	48	62	73	106	89
	12	52	43	38	17	18	15	20	6	2	0	0	0	1	19	45	80	80	56	46	59	49	127	100
16	10	28	33	30	17	25	16	14	4	0	0	0	3	20	54	169	89	38	47	36	53	81	93	90
	8	23	31	34	29	30	20	7	0	0	0	2	52	123	170	126	72	49	33	47	57	76	104	77
	14	22	24	47	46	23	5	0	0	0	2	29	104	402	572	99	40	39	50	57	45	41	84	62
	7	21	26	29	29	4	0	0	0	0	7	62	190	328	102	43	18	26	32	41	44	49	79	46
	4	14	10	7	3	0	0	0	0	11	39	111	133	54	31	22	19	14	17	23	34	61	62	35
	7	2	3	0	0	0	0	0	3	35	96	129	47	31	25	15	17	14	15	19	42	73	47	27
	2	2	2	0	0	0	0	1	32	114	66	51	15	16	22	16	14	21	28	48	80	55	24	14
23	3	0	0	0	0	0	1	5	22	45	17	16	13	11	10	12	11	12	13	36	30	18	7	15

FIGURE 16. Heat map representation of ODT matrix for Ω extracted from ground truth data at *YAKSU* station on April 1, 2016.

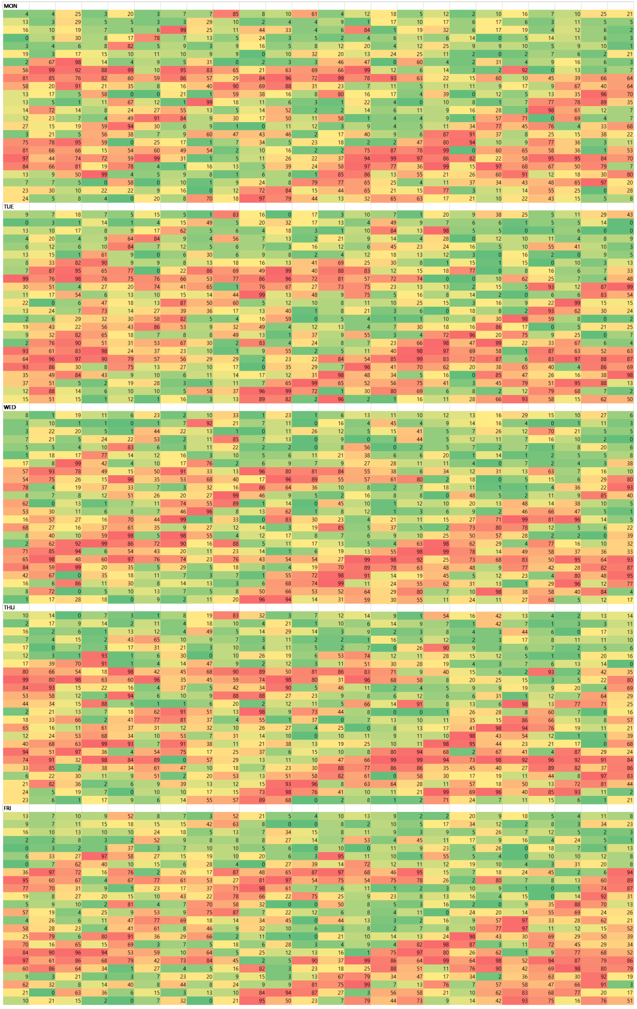


FIGURE 17. Heat map representation of Ω in terms of day of the week. From top to bottom, the five panels represent Monday to Friday, in order.

Figure 16 shows the heat map representation of the 24 × 24 matrix for Ω defined in Eq. 14. This figure is compared with Figure 9, the output of the simulation for *YAKSU* station in April. 1. 2016. Owing to lack of passengers (a single day data), $\omega_{ij} \in \Omega$ in the figure shows low granularity, but overall shape of the heatmap is similar to our simulation outputs at Figure 7.

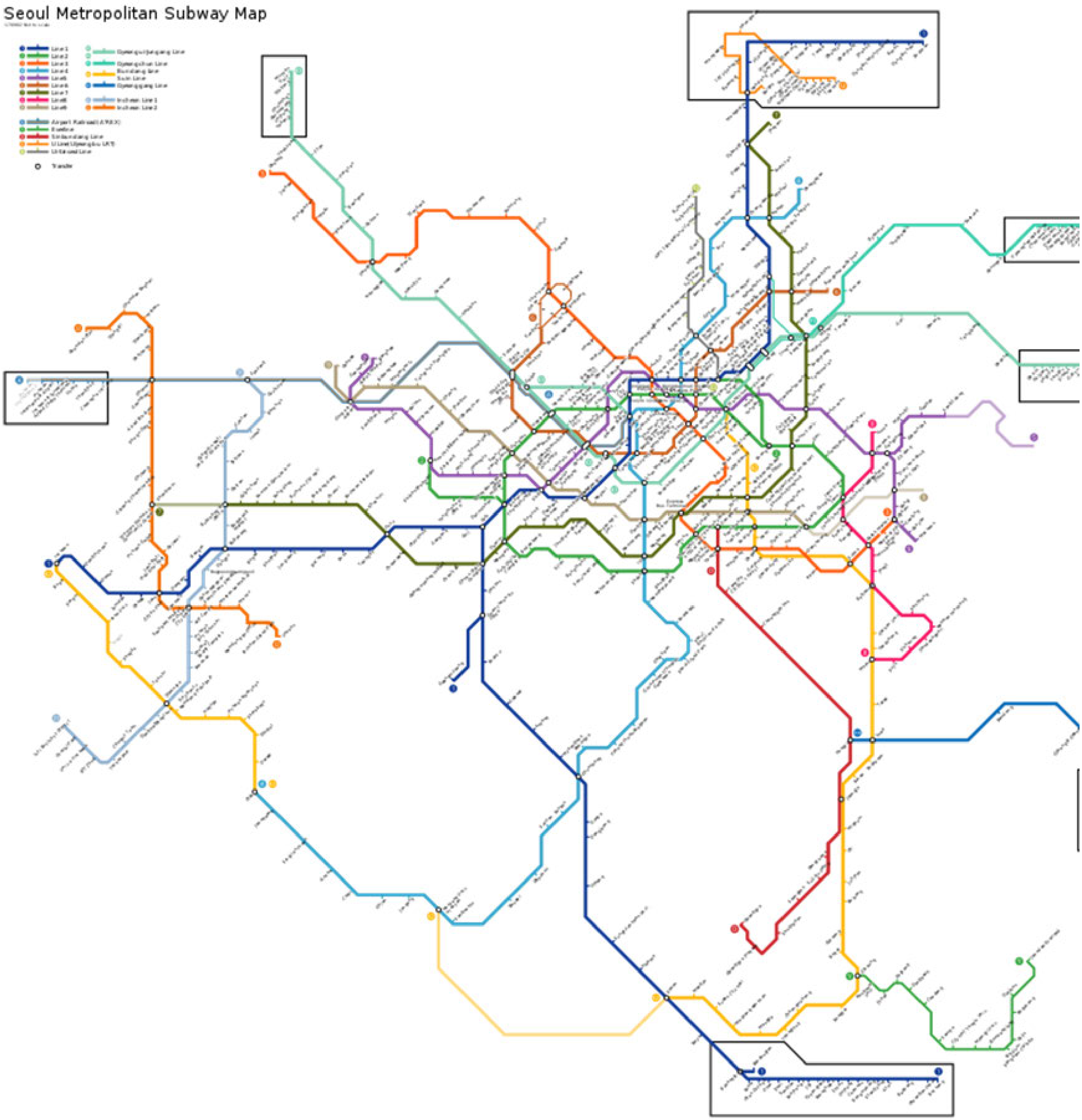


FIGURE 18.

VII. DISCUSSIONS AND CONCLUSION

Efficient and effective subway systems play a crucial role in the development of smart cities, as they are expected to constitute the major transportation systems of mega cities. With the rapid development of sensing technologies, subway systems generate big data, which is useful for the stable operation and effective management of subway stations. The unstructured and disorderly nature of big data, however, often obstructs effective information processing that can provide useful insights and identify meaningful patterns. To address the problem, this paper aimed at proposing an effective analytical approach based on the 'untraceable' big data collected from Seoul Metro. The untraceable data in this paper does not contain sufficient information to identify specific passengers' entry to and exit from stations and provides only the information regarding hourly cumulative entry and exit counts.

To extract hidden meaningful knowledge from the untraceable data, this paper adopted the inverse problem approach, used a fluid dynamics model and GA-based optimization solver, and generated heat maps that predict the lifestyles of residents. The results presented in this paper partially reconstructed the missing information (i.e., passenger ID) and estimated the general **ODT** patterns of passengers. The distribution of the elapsed time defined on an hourly basis taken until a passenger returns back to their station of origin was presented. The analysis results can help us to identify the lifestyles of passengers living in a specific area or around a specific subway station. The analysis of untraceable ridership data yielded the following findings:

1) Ridership data could be decomposed into a matrix Ω, indicating that hourly cumulative entry and exit numbers were decomposed into passenger enter and exit times.
2) a cell of Ω, ω_{ij}, could be interpreted as departure time at i-th hour and either **ODT** or **HDT** of passenger.
3) An earlier departure meant a longer stay away from home.
4) Weekday ridership had three clustered high probability regions, whereas weekend ridership had two regions.
5) Low ridership was observed between the 10th and 15th hours on weekdays.
6) The maximum probabilities of **ODT** were 2 hours on Sunday and 12 hours on weekdays.
7) Passengers stay away from their homes for a shorter period on Sunday than on a weekday.
8) The difference in ridership between weekdays and the weekend was mainly caused by passengers going to work in the morning hours before the 9th hour on weekdays.

It is noteworthy that the findings above resemble typical outputs for subway traffic flow research if the input data contains all the necessary information including individual passenger's transit records. However, this is not the case for the input data in this study. Thus, successful extraction of the missing information is essential.

Development of sensing technology for subway systems generates big data with huge potential, but we must overcome the complexity problems created by big data. Our analytical approaches present effective solutions to the problems associated with untraceable ridership subway data by helping to identify hidden patterns regarding the lifestyles of passengers. Our analytical approach and models contribute to solving the problem of extracting hidden knowledge from big data, especially when the data is missing critical information. Furthermore, the analytical models used to identify passenger lifestyles from the data can assist policy makers and subway operators in improving system performance, thus making smart cities smarter.

Finally, we present the limitation of the present study and our plans for future works below. In this study we presented a tangible case of applying a knowledge extraction paradigm using subway ridership in Seoul Metro, Korea, which lacked critical information, namely passenger IDs. Our work was partially successful in recovering the missing information (passenger ID) and could estimate the pattern of the **ODT** of passengers in s statistical convergence sense. A more detailed level of information reconstruction by developing more knowledge extraction methods would be our next objective. We presented the simulation results of ridership at the *YAKSU* station alone. By extending the target stations to all subway stations using the same model described herein, subway stations can be characterized in terms of passenger transit patterns. Improving the optimization solver specialized to big data is also an area of future work.

APPENDIX. SEOUL METROPOLITAN SUBWAY MAP

See Figure 18.

ACKNOWLEDGMENT

The authors would like to thank Yoon Jeong Lee, a government officer at Seoul Metropolitan Government, for providing the passenger transit record data of Seoul Metropolitan Subway.

REFERENCES

[1] M. Angelidou, "Smart cities: A conjuncture of four forces," *Cities*, vol. 47, pp. 95–106, Sep. 2015.

[2] A. Visvizi and M. D. Lytras, "Rescaling and refocusing smart cities research: From mega cities to smart villages," *J. Sci. Technol. Policy Manage.*, vol. 9, no. 2, pp. 134–145, Jul. 2018.

[3] M. Lytras and A. Visvizi, "Who uses smart city services and what to make of it: Toward interdisciplinary smart cities research," *Sustainability*, vol. 10, no. 6, p. 1998, Jun. 2018.

[4] A. Visvizi, M. D. Lytras, E. Damiani, and H. Mathkour, "Policy making for smart cities: Innovation and social inclusive economic growth for sustainability," *J. Sci. Technol. Policy Manage.*, vol. 9, no. 2, pp. 126–133, Jul. 2018.

[5] *The Seoul Research Database*. [Online]. Available: http://data.si.re.kr/statistics-seoul

[6] X. Wu and D. Theodoratos, "Homomorphic pattern mining from a single large data tree," *Data Sci. Eng.*, vol. 1, no. 4, pp. 203–218, Jan. 2017.

[7] C. Yadav, S. Wang, and M. Kumar, "Algorithm and approaches to handle large data—A survey," *Int. J. Comput. Sci. Netw.*, vol. 2, Jul. 2013.

[8] M. P. Derde and D. L. Massart, "Extraction of information from large data sets by pattern recognition," *Anal. Chem.*, vol. 313, no. 6, pp. 484–495, Jan. 1982.

[9] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, PA, USA: SIAM, 2005, pp. 1–9.

[10] M. Lytras, V. Raghavan, and E. Damiani, "Big data and data analytics research: From metaphors to value space for collective wisdom in human decision making and smart machines," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 1, pp. 1–10, 2017.

[11] M. Xiaolei, W. Yao-Jan, W. Yinhai, C. Feng, and L. Jianfeng, "Mining smart card data for transit riders' travel patterns," *Transp. Res. C, Emerg. Technol.*, vol. 36, pp. 1–12, Jul. 2013.

[12] C. Iliopoulou and K. Kepaptsoglou, "Combining ITS and optimization in public transportation planning: State of the art and future research paths," *Eur. Transp. Res. Rev.*, vol. 11, no. 1, Dec. 2019, Art. no. 27.

[13] R. Morello, S. C. Mukhopadhyay, Z. Liu, D. Slomovitz, and S. R. Samantaray, "Advances on sensing technologies for smart cities and power grids: A review," *IEEE Sensors J.*, vol. 17, no. 23, pp. 7596–7610, Dec. 2017.

[14] A. A. Guedes, J. C. Alvarenga, M. dos Santos Sgarbi Goulart, M. R. Y. Rodriguez, and C. P. Soares, "Smart cities: The main drivers for increasing the intelligence of cities," *Sustainability*, vol. 10, no. 9, p. 3121, 2018.

[15] N. Joshua, C. Daniel, and B. Matt, "How national governments can help smart cities succeed," Center Data Innov., Tech. Rep., 2017.

[16] M. Khan, M. Babar, S. H. Ahmed, S. C. Shah, and K. Han, "Smart city designing and planning based on big data analytics," *Sustain. Cities Soc.*, vol. 35, pp. 271–279, Nov. 2017.

[17] J. H. Lee, M. G. Hancock, and M.-C. Hu, "Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco," *Technol. Forecasting Social Change*, vol. 89, pp. 80–99, Nov. 2014.

[18] X. Yang, B. Ning, X. Li, and T. Tang, "A two-objective timetable optimization model in subway systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 1913–1921, Oct. 2014.

[19] P. Pan, H. Wang, L. Li, Y. Wang, and Y. Jin, "Peak-hour subway passenger flow forecasting: A tensor based approach," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3730–3735.

[20] M. Ni, Q. He, and J. Gao, "Forecasting the subway passenger flow under event occurrences with social media," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 6, pp. 1623–1632, Jun. 2017.

[21] Z. Wang and X. Cai, "Research on passenger flow prediction of beijing subway based on spatiotemporal correlation analysis," in *Proc. IEEE 4th Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, Apr. 2019, pp. 279–283.

[22] P. Wang, C. Wu, and X. Gao, "Research on subway passenger flow combination prediction model based on RBF neural networks and LSSVM," in *Proc. Chin. Control Decis. Conf. (CCDC)*, May 2016, pp. 6064–6068.

[23] Y. Sun, B. Leng, and W. Guan, "A novel wavelet-SVM short-time passenger flow prediction in beijing subway system," *Neurocomputing*, vol. 166, pp. 109–121, Oct. 2015.

[24] M. Lee and K. Sohn, "Inferring the route-use patterns of metro passengers based only on travel-time data within a Bayesian framework using a reversible-jump Markov chain Monte Carlo (MCMC) simulation," *Transp. Res. B, Methodol.*, vol. 81, pp. 1–17, Nov. 2015.

[25] L. Sun, Y. Lu, J. G. Jin, D.-H. Lee, and K. W. Axhausen, "An integrated Bayesian approach for passenger flow assignment in metro networks," *Transp. Res. C, Emerg. Technol.*, vol. 52, pp. 116–131, Mar. 2015.

[26] X. Yang, H. Dong, and X. Yao, "Passenger distribution modelling at the subway platform based on ant colony optimization algorithm," *Simul. Model. Pract. Theory*, vol. 77, pp. 228–244, Sep. 2017.

[27] S. Tao, D. Rohde, and J. Corcoran, "Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap," *J. Transp. Geogr.*, vol. 41, pp. 21–36, Dec. 2014.

[28] P. Liu, M. D. El Basha, Y. Li, Y. Xiao, P. C. Sanelli, and R. Fang, "Deep evolutionary networks with expedited genetic algorithms for medical image denoising," *Med. Image Anal.*, vol. 54, pp. 306–315, May 2019.

[29] T. Flognfeldt, "The tourist route system–models of travelling patterns," *Tourist Route Syst.-Models Travelling Patterns*, vol. 1, pp. 35–58, Oct. 2005.

[30] H. Li, R. Guensler, and J. Ogle, "Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1926, no. 1, pp. 162–170, Jan. 2005.

[31] T. Hodge, *Public Transportation's Role in Responding to Climate Change*. Darby, PA, USA: Diane Publishing, 2010.

[32] M. Karg and A. Kirsch, "A human morning routine dataset," in *Proc. Int. Conf. Auton. Agents Multi-Agent Syst.*, May 2014, pp. 1351–1352.

[33] N. Kazu, K. Katzsunao, and K. Ryuichi, "Empirical analysis of trip chaining behavior," *Transp. Res. Rec.*, vol. 203, no. 1203, pp. 48–59, 1988.

[34] B. Padmaja, V. V. R. Prasad, and K. V. N. Sunitha, "Use of reality mining dataset for human behavior analysis—A survey," in *Proc. Int. Conf. Inf. Syst. Eng. (ICISE)*, Apr. 2016, pp. 38–42.

[35] U. Blanke and B. Schiele, "Daily routine recognition through activity spotting," in *Proc. Int. Symp. Location Context Awareness*, May 2009, pp. 192–206.

[36] N. Banovic, T. Buzali, F. Chevalier, J. Mankoff, and A. K. Dey, "Modeling and understanding human routine behavior," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 248–260.

[37] C. Julien, G. Lefebvre, F. Ramparany, and J. L. Crowley, "Human activity recognition using place-based decision fusion in smart homes," in *Proc. Int. Interdiscipl. Conf. Modeling Using Context*, Jan. 2018, pp. 137–150.

[38] R. Carlos, J. C. Augusto, and D. Shapiro, "Ambient intelligence—The next step for artificial intelligence," *IEEE Intell. Syst.*, vol. 23, no. 2, pp. 15–18, Mar./Apr. 2008.

[39] R. Ortiz, J. Luis, A. Ghio, X. Parra, D. Anguita, J. Cabestany, and A. Catala, "Human activity and motion disorder recognition: Towards smarter interactive cognitive environments," in *Proc. ESANN*, 2013, pp. 1–10.

[40] E. Spissu, I. Meloni, and B. Sanjust, "Behavioral analysis of choice of daily route with data from global positioning system," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2230, no. 1, pp. 96–103, Jan. 2011.

[41] R. Attila, D. Stricker, and G. Hendeby, "Towards robust activity recognition for everyday life: Methods and evaluation," in *Proc. Int. Conf. Pervas. Comput. Technol. Healthcare*, 2013, pp. 25–32.

[42] A. Hande, H. Ertan, O. D. Incel, and C. Ersoy, "ARAS human activity datasets in multiple homes with multiple residents," in *Proc. Int. Conf. Pervas. Comput. Technol. Healthcare*, May 2013, pp. 232–235.

[43] L. Yu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *J. Geograph. Syst.*, vol. 14, no. 4, pp. 463–483, Oct. 2012.

[44] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[45] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the Internet of Things: Literature review and challenges," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 8, Aug. 2015, Art. no. 431047.

[46] R. David N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, Dec. 2011.

[47] Princeton University, Engineering School. (Dec. 2, 2013). *Forget the Needle, Consider the Haystack*. [Online]. Availiable: https://www.princeton.edu/news/2013/10/28/forget-needle-considerhaystack-uncovering-hidden-structures-massive-data

[48] Z. W. Geem, J. H. Kim, and G. V. Loganathan, "A new heuristic optimization algorithm: harmony search," *Simulation*, vol. 76, no. 2, pp. 60–68, Feb. 2001.

[49] Z. W. Geem, "Improved harmony search from ensemble of music players," in *Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Oct. 2006, pp. 86–93.

[50] M. Mahdavi, M. Fesanghary, and E. Damangir, "An improved harmony search algorithm for solving optimization problems," *Appl. Math. Comput.*, vol. 188, no. 2, pp. 1567–1579, May 2007.

[51] M. G. H. Omran and M. Mahdavi, "Global-best harmony search," *Appl. Math. Comput.*, vol. 198, no. 2, pp. 643–656, May 2008.

[52] Z. W. Geem and K.-B. Sim, "Parameter-setting-free harmony search algorithm," *Appl. Math. Comput.*, vol. 217, no. 8, pp. 3881–3889, Dec. 2010.

[53] L. Davis, *Handbook of Genetic Algorithms*. New York, NY, USA: Van Nostrand Reinhold, 1991.

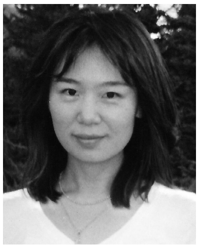
[54] A. H. Wright, "Genetic algorithms for real parameter optimization," *Found. Genet. Algorithms*, vol. 1, pp. 205–218, Jun. 1999.

[55] X. Jiang and C. Xiao, "Household energy demand management strategy based on operating power by genetic algorithm," *IEEE Access*, vol. 7, pp. 96414–96423, 2019.

[56] A. Iqbal, M. Meraj, M. Tariq, K. A. Lodi, A. I. Maswood, and S. Rahman, "Experimental investigation and comparative evaluation of standard level shifted multi-carrier modulation schemes with a constraint GA based SHE techniques for a seven-level PUC inverter," *IEEE Access*, vol. 7, pp. 100605–100617, 2019.

[57] M. Verotti, P. Di Giamberardino, N. P. Belfiore, and O. Giannini, "A genetic algorithm-based method for the mechanical characterization of biosamples using a MEMS microgripper: Numerical simulations," *J. Mech. Behav. Biomed. Mater.*, vol. 96, pp. 88–95, Aug. 2019.

[58] *Wolfram Wiki*. [Online]. Available: http://mathworld.wolfram.com/DiracDeltaFunction.html



HYUNKYUNG SHIN received the Ph.D. degree in applied mathematics and statistics from Stony Brook University, NY, USA, in 2002. She has been an Associate Professor with the Department of Financial Mathematics (mathematical science), Gachon University, South Korea, since 2007. She currently serves as the Director of the Institute for Intelligent Transportation Technology. As a chief researcher of a government research project, she developed an automation module of an NLP for analyzing Alzheimer's disease. She also developed a text recognition software through the cooperation of Korean IT industry. Her research interests include image processing, machine learning, and mathematical education.