An Analysis on Movement Patterns between Zones Using Smart Card Data in Subway Networks

Kwanho Kim[1], Kyuhyup Oh[2], Yeong Kyu Lee[3], SungHo Kim[3], and Jae-Yoon Jung[2†]

[1] Department of Industrial and Management Engineering, Incheon National University, Incheon, South Korea

[2] Department of Industrial and Management Systems Engineering, Kyung Hee University, Yongin, South Korea

[3] Seoul Metropolitan Rapid Transit Corporation, Seoul, South Korea

[†] Corresponding author

Abstract. Identifying zones and movement patterns of people is crucial to understanding adjacent regions and the relationship in urban areas. Most previous studies addressed zones or movement patterns separately without analysing simultaneously the two issues. In this paper, we propose an integrated approach to discover directly both zones and movement patterns among the zones, referred to as *movement patterns between zones* (MZPs), from historical boarding behaviours of passengers in subway networks by using an agglomerative clustering method. In addition, evaluation measures of MZPs are suggested in terms of coverage and accuracy. The effectiveness of the proposed approach is finally demonstrated through a real-world dataset obtained from smart cards on a subway network in Seoul, Korea.

Keywords: zone analysis; movement pattern analysis; boarding behaviour; movement clustering; smart card data; subway networks

1. Introduction

Understanding geographically adjacent regions based on the movements of passengers is essential to facilitate various location-based activities such as residential area recommendation and urban planning (Antikainen, 2005; Yuan, Zheng and Xie, 2012). Here, a set of such geographically adjacent stations is called a *zone* (Fusco and Caglioni, 2011). The notion of zone is reasonable since adjacent stations are often

significantly coherent with each other in terms of function due to their geographical proximities in most urban areas. In this research, a *movement pattern between zones* (MZP) is defined as a pair of zones, an origin and a destination, that are strongly related with each other in terms of people's movements. MZPs are useful not only to articulate urban development strategies according to discovered zones, but also to improve passenger experiences in public transportation systems through re-scheduling resources and planning additional transportation methods. Moreover, based on the estimated flows of people between zones, it is possible to design the advanced advertisements that take transit flow into consideration (Blythe, 2004; Trépanier and Morency, 2010). By utilizing discovered MZPs, advertisers are provided with segmented customers with respect to movement behaviours, enabling more precisely targeted advertising.

Due to the dynamic nature of people's movements, it is challenging to identify zones and their relationships, which are continuously changing according to the daily experiences of people, compared to relatively stationary city development plans (Bagchi and White, 2005). However, with the recent advent of electronic-card payment systems (EPSs) which automatically charge a passenger with a transit fee by touching an electronic-chip-embedded card, called a *smart card*, movement behaviours of people on a transportation network such as subway and bus have become easier to record and investigate (Blythe, 2004). Since each smart card is associated to a unique identifier, an EPS is able to track the origin and the destination stations for each movement from the records such as stations, line, and transit time into an origin-destination (OD) database. Currently, EPSs are widely available in many modern transportation networks across the world such as UK, France, Finland, Italy, China, and Korea (Pelletier, Trépaniera and Morency, 2011).

Many studies on the analysis of OD datasets have conducted to discover movement patterns among specific regions in urban areas (Bagchi and White, 2005; Joh *et al*., 2001; Lee and Mark, 2011; Lee and Park, 2005; Srinivasan and Ferreira, 2003) and to characterize the patterns of movement sequences (Chu and Chapleau, 2010; Hoffman *et al*., 2009; Ma *et al*., 2013). Particularly, some research mainly focused on analysing movements based on specific factors such as commutes (Bhat, 2001; Fusco and Caglioni, 2011; Konjar *et al*., 2010), travel time (Jang, 2010; Morency *et al*., 2006; Zhao *et al*., 2013), individual behaviours (Liu *et al*., 2009), and trajectories (Ghasemzadeh *et al*., 2014). Most of them assume the predefined regional partitions to understand the people's movement trends that mainly represent from and to which

places people usually move. Some of the authors develop pattern discovery algorithms and suggest measures for quantifying the effectiveness of movement patterns.

In the meantime, there are some methods for predicting the amount of people's movements between points of regions (Chu and Chapleau, 2008; Park *et al*., 2008; Munizaga *et al*., 2010; Munizaga and Palma, 2012). They have estimated the future movements of people between regions based on the past observations between the regions. Furthermore, a few movement estimation approaches have been suggested to enhance resource scheduling (Bagchi and White, 2004; Trépanier *et al*., 2009). Along with the research focused on movement patterns of people, a few methods for identifying zones have been also proposed (Fusco and Caglioni, 2011; Karlsson, 2007; Konjar *et al*., 2010; Martin, 2003; Moreno-Regidor *et al*., 2012). By applying clustering analysis, they have focused on discovering coherent functional regions based on movement behaviours of people.

Although the previous research successfully identified either zones or movement patterns, they still have limitations to directly capture MZPs from movement observations. They concentrate on addressing zone identification while other studies mainly focus on unveiling movement patterns of people. To obtain MZPs through the previous approaches, combining multiple existing methods is required, which is difficult due to their heterogeneity. Moreover, such combined models cannot guarantee satisfactory results by taking into account the balance between zone identification and movement pattern discovery.

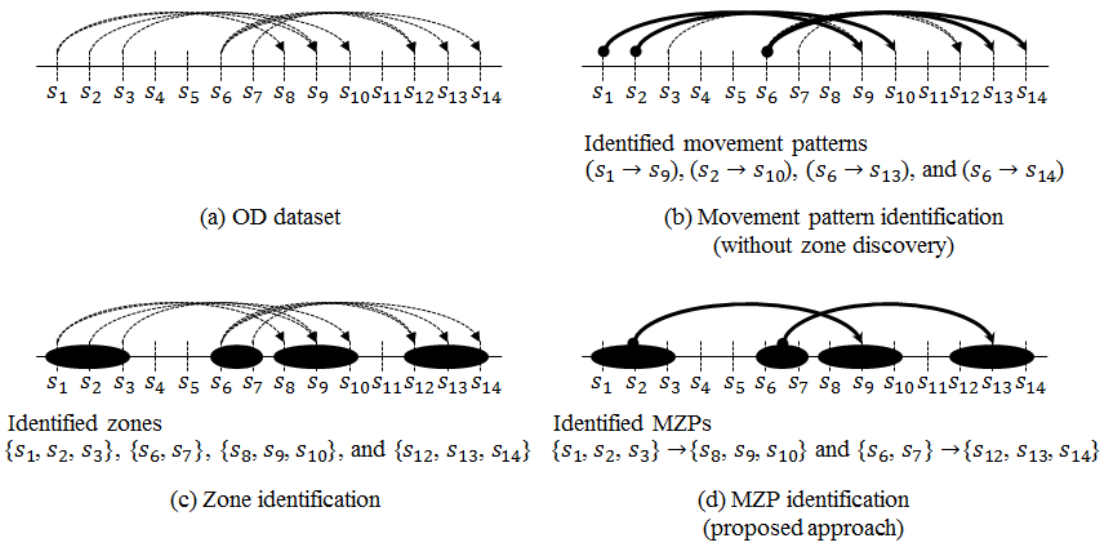

Figure 1. Comparison of movement pattern and zone identification approaches.

Figure 1 illustrates the comparison of our proposed approach to the existing ones. For a given OD dataset shown in Figure 1 (a), it is assumed that there exist possible zones such as $\{s_1, s_2, s_3\}$, $\{s_8, s_9, s_{10}\}$, $\{s_6, s_7\}$, and $\{s_{12}, s_{13}, s_{14}\}$, where the first zone is closely related to the third zone whereas the second zone is related to the last zone. Figure 2 (b) shows that an approach that only focuses on movement patterns between points of regions is restricted to identify movement patterns such as $(s_1 \longrightarrow s_9)$, $(s_2 \longrightarrow s_{10})$, $(s_6 \longrightarrow s_{13})$, and $(s_6 \longrightarrow s_{14})$ without suggesting zones. On the other hand, Figure 2 (c) represents an approach that is mainly modelled to identify zones fails to discover the hidden relations between zones. On the contrary, the MZP identification approach is capable to directly unveil both zones and the relations among them, as shown in Figure 1 (d).

In this research, we develop an integrated approach to identifying MZPs from historical movement behaviours of people on subway networks. The objective of our approach is to simultaneously discover zones by combining adjacent regions and hidden movement patterns between zones. Through addressing the issue , the proposed approach is designed to yield well-balanced results in zoning and patterning movements. Therefore, the approach is beneficial in resolving transportation and urban development issues that often require considering not only zone discovery but also the movement patterns between zones, at the same time.

Specifically, an MZP identification algorithm is developed, which iteratively identifies MZPs from individual observations for a given OD dataset without predefined zone partitions. At each iteration step, the proposed algorithm attempts to search for a better MZP by combining two MZPs into a single one based on their adjacent proximity. Identified MZPs are then evaluated by using two measures, coverage and accuracy, which are taking into account the number of movements and the dependency between zones, respectively. As a new MZP is identified by combining two existing ones, the MZP becomes stronger in terms of coverage compared to the existing ones, while it is likely to become less strong in terms of accuracy due to the increments of its zone length. Therefore, to consider the trade-off caused by MZP merging, the effectiveness of MZPs are finally evaluated in terms of both coverage and accuracy by introducing a combined measure after enumerating possible MZPs from an OD dataset.

By using the proposed algorithm, MZP identification tasks according to iteration steps are visually presented, and top ranked MZPs are analysed based on a real-world

OD dataset obtained from a subway network in Seoul, Korea. Moreover, the distributions of MZPs according to coverage and accuracy are suggested, showing that only a few highly effective MZPs exist while the majority of MZPs are negligible to explain the entire dataset. In addition, by examining identified MZPs with respect to commuting behaviours of people, we found some MZPs that were more suitable to address a particular type of commuting behaviours while the others yielded quite low performances regardless of commuting behaviours.

This paper is organized into the following sections. First, the proposed approach to MZP analysis is developed in Section 2. In Section 3, we suggest three evaluation measures to quantify the effectiveness of MZPs with respect to coverage and accuracy. In Section 4, the experimental results obtained by the proposed approach are demonstrated using a real-world OD dataset from a subway network. Finally, we discuss relevant findings in Section 5 and complete the analysis.

2. Movement pattern analysis

2.1. OD dataset

The attributes of an OD dataset differ according to EPSs caused by different aspects of operation, maintenance, and management policies. There originally existed 24 distinct attributes in the OD dataset used for this research such as smart card identifiers, associated stations, transit times, lines, passenger types, train identifiers, and transfer information. Among them, only four attributes, origin station, destination station, time, and line. The attributes are directly related with from where, to where, and when a person moves, as shown in Table 1. The other attributes do not necessarily depend on people's movements in a subway network. In this research, it is assumed that each movement is associated to both its origin and destination stations. The origin and destination information can be often obtained in EPSs especially in subway networks since the EPSs calculate transit fees of passengers based on their movement distances.

Table 1. OD record attributes used in this research.

No	Field	Description	Data type
1	Origin station	Station number (from where)	Numeric
2	Destination station	Station number (to where)	Numeric
3	Time	Time to ride at origin station (when)	Date time
4	Lines of origin and destination stations	Subway lines on which origin and destination stations exist	String

2.2. Research framework

The proposed approach is divided into two major phases, data acquisition and filtering and MZP identification. Figure 2 illustrates the details of involved tasks and information flows with examples of OD records and identified MZPs. When passengers touch their smart cards on card readers to pay their fees, their information containing origin and destination stations are recorded in an OD database. In the data acquisition and filtering phase, such OD records are retrieved for the experiments. In addition, to analyse MZPs according to movement time, multiple filtered OD datasets are constructed according to time slots. The OD datasets are used for the input of the proposed MZP identification algorithm.

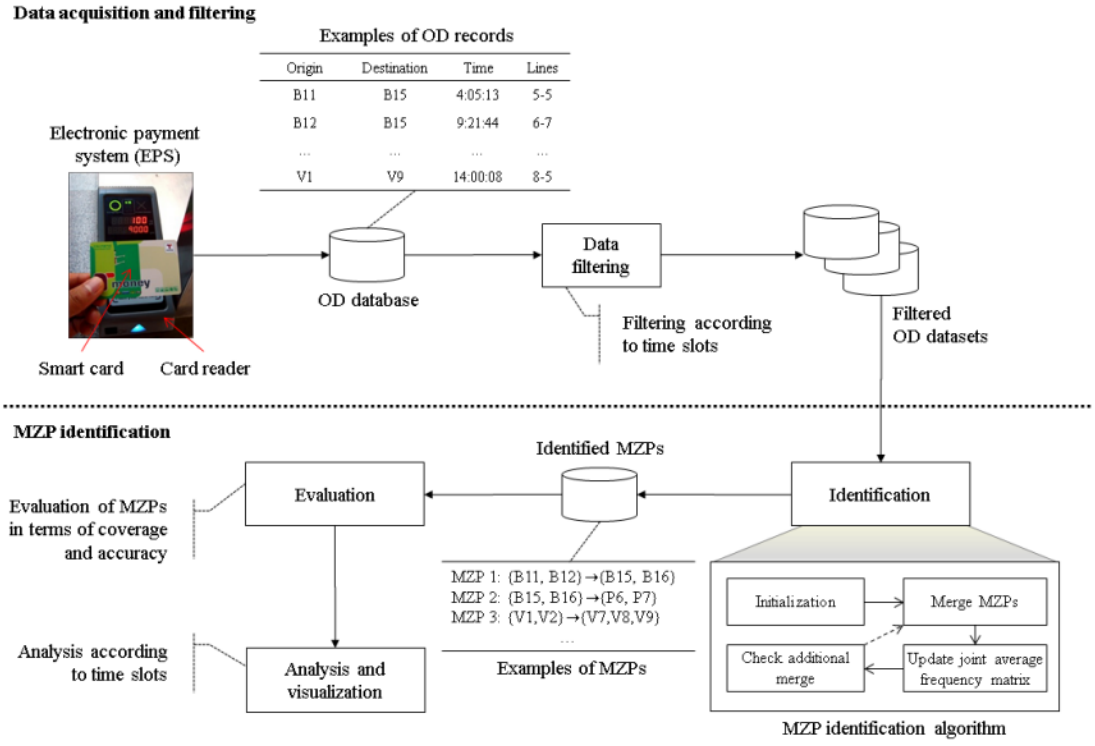

Figure 2. Overall research framework.

In the MZP identification phase, MZPs are identified from an OD dataset by utilizing the proposed algorithm. Based on the observed movements in the dataset, the proposed algorithm attempts to search for better MZPs through iteratively combining OD records. Since the adjacency between stations and the directions of observed movements are investigated, the outputs of the algorithm are guaranteed to be MZPs without further manipulations. The details of the proposed algorithm are presented in Section 2.3. All of the identified MZPs from the OD dataset are then evaluated, which is described in Section 3. Because this study mainly addresses on how MZPs are effectively identified and reasonably evaluated, we focus on the second part of the research framework throughout the remainder of this article.

2.3. Discovery of movement pattern between zones (*MZP*)

The proposed MZP discovery approach finds hidden MZPs through iteratively merging MZPs given a set of stations, *S*, and a set of observed movements, *V*. Figure 3 shows how MZPs are iteratively merged for given 12 observed distinct movements, denoted as $p_1, \dots, p_{12}$, and 7 stations, denoted as $s_1, \dots, s_7$. In the figure, two stations with subsequent indexes are assumed to be adjacent. The origin and destination stations of initial MZPs are shown in Figure 3 (a). Suppose that a pair of MZPs, $(p_7,\ p_8)$, which yields the most frequent movements on the average, is selected to be merged into a single MZP, denoted as p_A, as depicted in Figure 3 (b). Subsequently, three MZPs, p_B, p_C, and p_D, are identified by merging each pairs of MZPs, $(p_1,\ p_2)$, $(p_5,\ p_6)$, and $(p_{11},\ p_{12})$ through three iterations, as shown in Figure 3 (c). Finally, four MZPs, $p_{A'}$, $p_{B'}$, p_C, and $p_{D'}$, are identified after four more iterations, as presented in Figure 3 (d).

The underlying rationale of this approach is based on the agglomerative clustering analysis that iteratively combines two similar instances in a hierarchical manner. However, the proposed approach differs from the conventional agglomerative clustering method. Our approach conducts MZP merges that preserve the direction of each movement and the adjacency of stations in a zone rather than simply combining the nearest two observations. Moreover, while the conventional method is used for partitioning entire observations into subsets based only on their similarities, our approach aims to allow practitioners to identify several significant MZPs for better understanding of representative movements between zones.

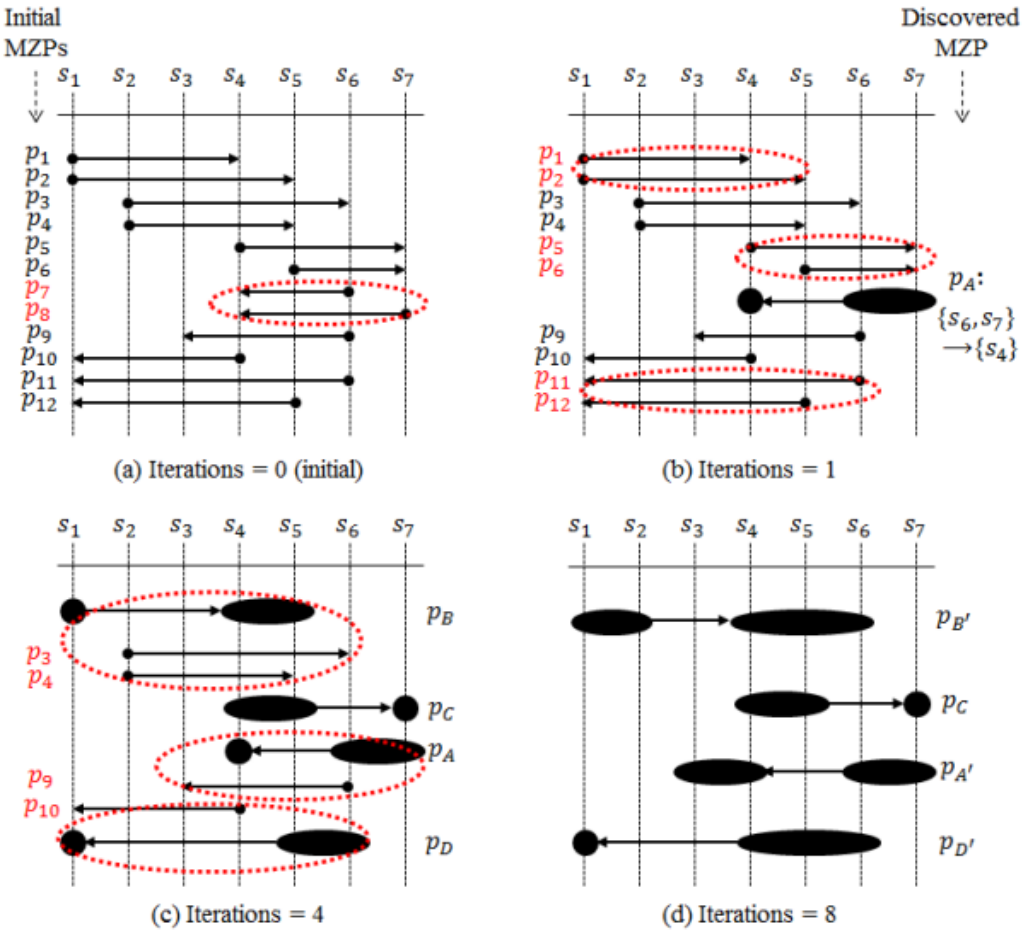

Figure 3. MZP examples iteratively identified from observed movements.

In this research, given a set of N stations, $S = \{s_1, \dots, s_N\}$, a set of M observed movements is denoted as $V = \{v_1, \dots, v_M\}$, where the m-th observed movement, $v_m = s_o \longrightarrow s_d$, is represented by a directional relation from an origin station, $s_o \in S$, to a destination station, $s_d \in S$. The i-th discovered MZP, $p_i = O_i \longrightarrow D_i$, represents the movement pattern from its origin zone, O_i, to its destination zone, D_i, where O_i and D_i are sets of adjacent stations in S. Moreover, it is said that movement $v_m = s_o \longrightarrow s_d$ is *covered* by MZP $p_i = O_i \longrightarrow D_i$, if $s_o \in O_i$ and $s_d \in D_i$. If all the movements covered by MZP p_i can be covered by MZP p_j, we can say that p_i is a subset of p_j. And, two MZPs, $p_i = O_i \longrightarrow D_i$ and $p_j = O_j \longrightarrow D_j$, are defined to be *adjacent* if there exist two pairs of stations, $(s_{oi} \in O_i,\ s_{oj} \in O_j)$ and $(s_{di} \in D_i,\ s_{dj} \in D_j)$, such that $d(s_{oi}, s_{oj}) \leq 1$ and $d(s_{di}, s_{dj}) \leq 1$, where $d(s_\cdot, s_\cdot)$ represents the distance between stations. The distance between stations is simplified to be one in this research since the geographical

distances between two adjacent stations in urban areas are generally similar to each other in a subway network. For instance, in the subway network considered in this research, the distances of 83% of adjacent stations range from 0.6 to 1.2 kilometres. In addition, the adjacencies between stations on different lines are considered, as well as those between stations on the same line, to be able to identify MZPs across different lines.

Based on the notations defined above, we define the average frequency measure of an MZP in an OD dataset to evaluate how many OD records in the dataset are covered by the MZP. Here, the average frequency of MZP $p_i = O_i \longrightarrow D_i$ represents the average number of observed movements from a station in O_i to a station in D_i. The average frequency of MZP p_i in an OD dataset is defined as:

$$\rho_i = \rho(p_i) = \frac{freq(O_i \to D_i)}{|O_i| \cdot |D_i|} \tag{1}$$

where $freq(O_i \to D_i)$ represents the total number of observed movements from stations in O_i to stations in D_i, and $|\cdot|$ is the size of the given set.

Then, the definition of the average frequency of an MZP is extended into the joint average frequency of two MZPs in order to process the merge of two MZPs. The joint average frequency of two MZPs, $p_i = O_i \longrightarrow D_i$ and $p_j = O_j \longrightarrow D_j$, is the average number of observed movements from one of their combined origin stations, $O_i \cup O_j$, to one of their combined destination stations, $D_i \cup D_j$. In the proposed algorithm, we assume that the joint average frequency can be achieved only if two MZPs are adjacent (i.e. can be merged). The joint average frequency of two MZPs, p_i and p_j, is defined as:

$$\rho_{i,j} = \rho(p_i, p_j) = \begin{cases} \dfrac{freq(O_i \cup O_j \to D_i \cup D_j)}{|O_i \cup O_j| \cdot |D_i \cup D_j|} & \text{if } p_i \text{ and } p_j \text{ are adjacent,} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

By utilizing Equation (2), the joint average frequency matrix that presents the joint average frequency values for all the pairs of MZPs is constructed or updated at each iteration step of MZP identification tasks. We note that $\rho_{i,j} \neq \rho_{i,j}$ by definition. For given *T* initial MZPs, where *T* is the number of distinct observed movements in a given OD dataset, the sets of MZPs at iterations, $1, \dots, K$, are respectively denoted as $P^{(1)} = \{p_1, \dots, p_T\}, \dots, P^{(K)} = \{p_1, \dots, p_{R_K}\}$, where *K* is the number of iterations and R_K

represents the number of the remaining MZPs at the *K*-th iteration. In details, $R_k = T + 1 - k - \sum_{k'=1}^{k-1} \Delta(k')$, where $\Delta(k')$ is a function that returns the number of MZPs which are a subset of the best merged MZP at the k'-th iteration. Therefore, the joint average frequency matrices for respective iterations, $C^{(1)}, \ldots, C^{(K)}$, are organized as:

$$C^{(1)} = \begin{pmatrix} \rho_{1,1}^{(1)} & \cdots & \rho_{1,T}^{(1)} \\ \vdots & \ddots & \vdots \\ \rho_{T,1}^{(1)} & \cdots & \rho_{T,T}^{(1)} \end{pmatrix}, \ldots, C^{(K)} = \begin{pmatrix} \rho_{1,1}^{(K)} & \cdots & \rho_{1,R_K}^{(K)} \\ \vdots & \ddots & \vdots \\ \rho_{R_K,1}^{(K)} & \cdots & \rho_{R_K,R_K}^{(K)} \end{pmatrix} \quad (3)$$

The MZP identification algorithm is presented to search for the best current merge of two MZPs at each iteration on the basis of the joint average frequencies $\rho_{i,j}$ and their matrix C, as shown in Figure 4. The algorithm sets each station as a zone at the initial iteration (line 2 in the figure), and it builds *T* MZPs by regarding each distinct observed movement as an MZP for the next iteration (line 3). Moreover, based on the initial MZPs, the initial joint average frequency matrix, $C^{(1)}$, is calculated using Equations (2) and (3) (line 4). At the *k*-th iteration, the algorithm finds the maximum value in $C^{(k)}$, denoted as $\rho^{*(k)} = \max\left(\rho_{i,j}^{(k)}\right)$ for $i, j = 1, \ldots, R_k$ $(i \neq j)$, and merges the two MZPs associated to $\rho^{*(k)}$ (line 6). At that time, if there are other MZPs which are subsets of the new merged MZP, then all the subset MZPs are also merged to new one (line 7). The number of the subsets is $\Delta(k')$. At each iteration, the algorithm also updates the joint average frequency matrix for the next step (line 8). MZP merges are repeatedly processed until there are no remaining MZP to merge or the maximum value in the joint average frequency matrix is less than threshold θ (line 9), and the remaining MZPs are finally returned (line 10).

1: **Initialize**
2: Set each station in *S* as a zone.
3: Build *T* MZPs for the distinct observed movement in *V*.
4: Calculate the joint average frequency matrix $C^{(1)}$ based on $\rho_{i,j}^{(1)}$ for $i, j = 1, \ldots, T$.
5: **Repeat**
6: Merge two MZPs, p_u and p_v, s.t. $\rho_{u,v}^{(k)} = \rho^{*(k)} = \max\left(\rho_{i,j}^{(k)}\right)$ for $i, j = 1, \ldots, R_k$ $(i \neq j)$.
7: If any other MZP is a subset of the new merged MZP, then also merge it to the new one.
8: Update the joint average frequency matrix $C^{(k)}$.
9: **Until** There is no MZP to merge or $\rho^{*(k)}$ is less than threshold θ.
10: **Return** The remaining MZPs.

Figure 4. MZP identification algorithm.

The proposed algorithm guarantees that further iterations are meaningless in terms of average frequency since the largest value in the joint average frequency matrix is decreasing as iterations are performed. Briefly, the joint average frequency of any two MZPs, shown in Equation (2), is not increased after merging due to $\rho_{i,j}^{(k)} \leq \max\left(\rho_{i,i}^{(k)}, \rho_{j,j}^{(k)}\right)$, for $u, v = 1, \ldots, R_k$. This means that, when the algorithm stops, it yields MZPs associated to higher values of joint average frequencies compared to ones that have not yet been identified.

The proposed algorithm can result in a significantly reduced computational cost and time compared to the exhaustive approaches that enumerate all possible MZPs to search for optimal MZPs especially when dealing with the subway networks which consist of a large number of stations. In the initial step, for the given *T* initial MZPs, which is the distinct observed movements in an OD dataset, the computational complexity of the proposed algorithm is $O(T^2)$ in constructing the joint average frequency matrix of *T* by *T* dimensions. During the remaining MZP discovery steps, on the other hand, the complexity becomes $O(T)$ in the worst cases since equal or more than one MZP is removed at each iteration step of MZP merges.

3. MZP evaluation

We suggest three measures to quantify the effectiveness of an MZP in terms of *coverage* and *accuracy*. The measures are extended from the previously developed guidelines for evaluating rules that represent the relation between two sets of features (He *et al*., 2012). Particularly, we adopt support, lift, and cosine metrics, which aim to evaluate a rule in terms of the number of covered instances by the rule, the correlation between sets of features in the rule, and the similarity between sets of features in the rule, respectively. Since each MZP can be regarded as a rule that explains the directed movements of people between zones, such metrics are suitable for evaluating MZPs in our context. Unlike the existing measures that assume simultaneous observations across all stations in a zone, we modified them to fit our specific problem through taking into account movements between stations each of which is in its origin or destination zones.

First, we measure the *coverage* of an MZP based on the frequency of the observed movements covered by the MZP. The coverage of MZP p_i for an OD dataset, called *v*-value, is defined as:

$$v(p_i) = \Pr(O_i \longrightarrow D_i) = \frac{freq(O_i \rightarrow D_i)}{M} \tag{4}$$

where $\Pr(O_i \longrightarrow D_i)$ is the probability that a movement from an origin station in O_i to a destination station in D_i is observed in an OD dataset, $freq(O_i \rightarrow D_i)$ represents the number of movements covered by $p_i = O_i \longrightarrow D_i$ in the dataset, and *M* is the number of all the observed movements in the dataset.

Second, we calculate the *accuracy* of an MZP by examining the dependency between its origin and destination zones. While the coverage of an MZP is calculated based on how many movements are conformant to the MZP, the accuracy of an MZP focuses on how much dependent its origin zone is on its corresponding destination zone. For instance, an MZP is valuable if people who ride on one of the stations in its origin zone frequently alight on one of the stations in its destination zone although its observed movement frequency is not very high. The accuracy of MZP p_i for an OD dataset, called a-value, is calculated as:

$$a(p_i) = \frac{\Pr(O. \longrightarrow D_i \mid O_i \longrightarrow D.)}{\Pr(O. \longrightarrow D_i)} = \frac{\Pr(O_i \longrightarrow D. \mid O. \longrightarrow D_i)}{\Pr(O_i \longrightarrow D.)}$$

$$= \frac{\Pr(O_i \longrightarrow D_i)}{\Pr(O_i \longrightarrow D.)\ \Pr(O. \longrightarrow D_i)} = \frac{M\ freq(O_i \rightarrow D_i)}{freq(O_i \rightarrow D.)\ freq(O. \longrightarrow D_i)} \tag{5}$$

where $\Pr(O. \longrightarrow D_i)$ and $\Pr(O_i \longrightarrow D.)$ are the probabilities that the destination station of an observed movement in an OD dataset belongs to D_i and the origin station of an observed movement belongs to O_i, respectively, $\Pr(O. \longrightarrow D_i \mid O_i \longrightarrow D.)$ represents the probability that the destination station of an observed movement in an OD dataset belongs to D_i, given that its origin station belongs to O_i, and $\Pr(O_i \longrightarrow D. \mid O. \longrightarrow D_i)$ is the probability that the origin station of an observed movement in an OD dataset belongs to O_i, given that its destination station belongs to D_i.

In Equation (5), for an OD dataset, $a(p_i)$ such that $p_i = O_i \longrightarrow D_i$ becomes higher than one if its two zones, O_i and D_i, are positively correlated while it gets closer to zero if the two zones are negatively correlated. In case that the origin and the destination zones are uncorrelated, $a(p_i)$ has near one since $\Pr(O. \longrightarrow D_i \mid O_i \longrightarrow D.) = \Pr(O. \longrightarrow D_i)$ and $\Pr(O_i \longrightarrow D. \mid O. \longrightarrow D_i) = \Pr(O_i \longrightarrow D.)$. Therefore, as more people

who ride on one of the stations in O_i alight on one of the stations in D_i, and vice versa, the accuracy value of MZP p_i becomes higher.

Finally, we additionally propose a measure that considers the trade-off between coverage and accuracy. Specifically, as MZPs are merged, the coverage of an MZP increases while its accuracy is likely to decrease. It is because the coverage measures MZP based on the number of movements covered by the MZP while the accuracy considers the dependency between two specific zones of the MZP. The trade-off can be ascertained when we consider MZPs consisting of a single specific zone closely related to other multiple zones. In such MZPs, it is highly possible that their *v*-values become quite high whereas their *a*-values become low (Tan *et al*., 2004). Therefore, through combining coverage and accuracy described in Equations (4) and (5), respectively, the *combined* measure for MZP $p_i = O_i \longrightarrow D_i$, called *c*-value, given an OD dataset, is defined as:

$$c(p_i) = \sqrt{v(p_i)a(p_i)} = \sqrt{\frac{freq(O_i \to D_i)}{M}\frac{M\ freq(O_i \to D_i)}{freq(O_i \to D.)\ freq(O. \to D_i)}} = \frac{freq(O_i \to D_i)}{\sqrt{freq(O_i \to D.)\ freq(O. \to D_i)}} \quad (6)$$

4. Experiment results

4.1. Dataset description

We collected an OD dataset for five consecutive days, from 18 Jun, 2012 (Monday) to 22 Jun, 2012 (Friday), from an EPS for a subway network in Seoul, Korea. The subway network was associated with 148 stations across four lines, Violet (V), Brown (B), Olive (O), and Pink (P), and the numbers of stations on each line were 51, 38, 42, and 17, respectively (See Table 3 in Appendix for the details of stations). 5,405,736 movements were obtained to be the OD dataset for the experiments, and the time slots in a day, ranged from 5:00 to 24:00. The proportions of the observed movements in lines V, B, O, and P were 35%, 17%, 38%, and 10%, respectively. For the purpose of noise reduction, we removed observed movements associated with a pair of an origin and a destination stations if the number of observed movements between the stations was less than a given value, called minimum movement value, which was

one hundred in the experiments. Finally, 4,763,823 observed movements were used in our experiments, which were visualized on a map of Seoul, as shown in Figure 5.

We note that, as the minimum movement value becomes larger, the distances between zones in discovered MZPs tend to be closer, while the smaller minimum movement value is likely to yield the larger distances between zones in the discovered MZPs. It is because the larger minimum movement values imply removing more movements associated with distant zones rather than close zones. We empirically found that minimum movement values ranging from 50 to 200 were suitable to uncover the overall movement patterns in the dataset we considered, and there was no significant difference according to the minimum movement values in the considered range. However, in order to adjust the minimum movement value in a more effective way, further methods to sophisticatedly estimate the appropriate values according to a given dataset.

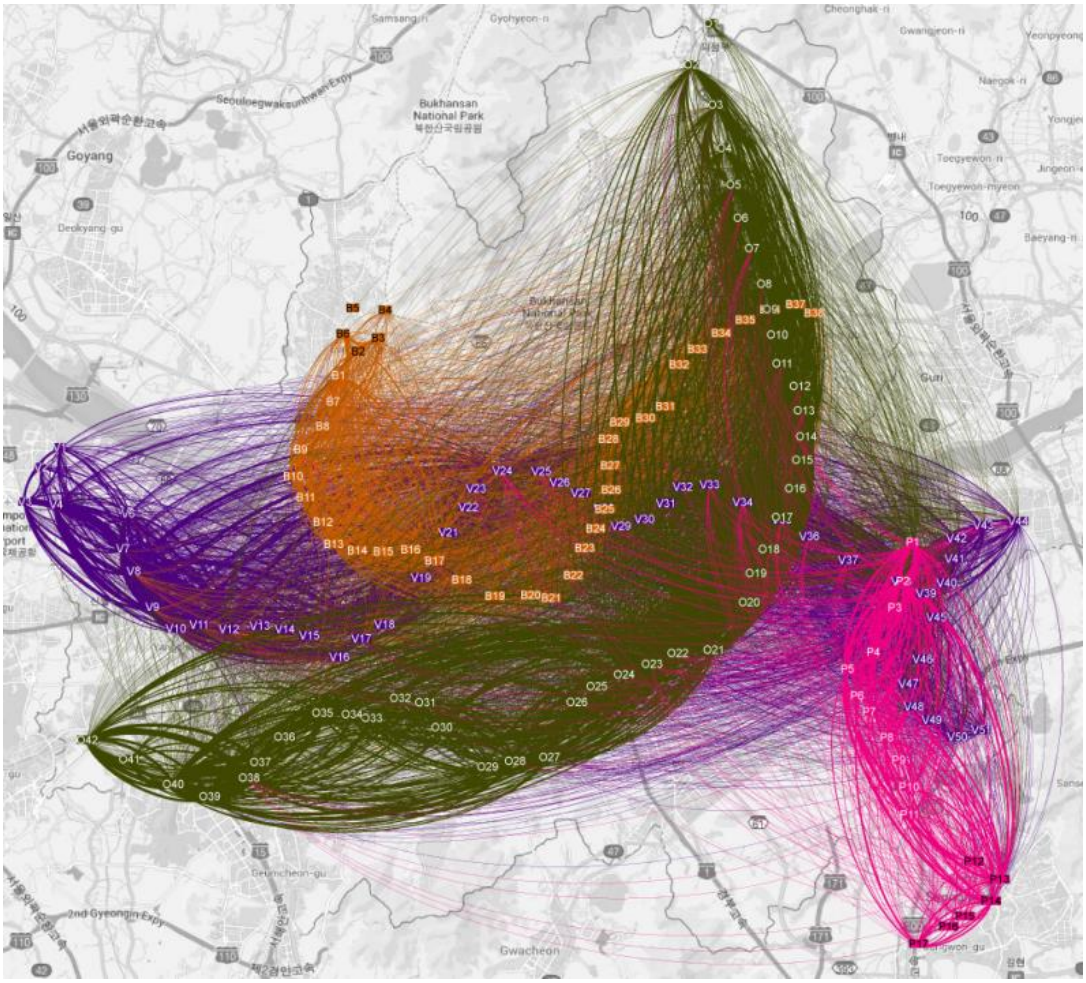

Figure 5. Visualization of the OD dataset.

Figure 6 depicts the overall distributions of the collected dataset in terms of movement distances and time slots, implying that people tend to move a short distance and people's movements are intensively performed at some specific time slots. In details, Figure 6 (a) shows that about half of the observed movements passed less than 9 stations, and only 1% of the observed movements passed more than 38 stations. The most frequently observed movement distance was 3 with a proportion of 7.304%, and the largest movement distance was 61 with a frequency of 1. The mean and the variance of the movement distances were 10.037 and 63.858, respectively, and their median value was 8.

Figure 6 (b) illustrates the ratios of observed movements according to time slots. A time slot is regarded as rush hours if the appearance ratio of movements observed at the time slot larger than the average appearance ratio of movements at a time slot across the entire time slots considered, while the rest of time slots are considered idle hours. Since the average appearance ratio of observed movements at a time slot was 0.05 in our dataset, and 7:00 to 10:00 and 17:00 to 20:00, were founded as rush hours. The number of observed movements in the rush hours was 2,130,278, with a portion of 44.718%, indicating, on the average, 8.944% of the entire movements were observed in a single time slot in the rush hours.

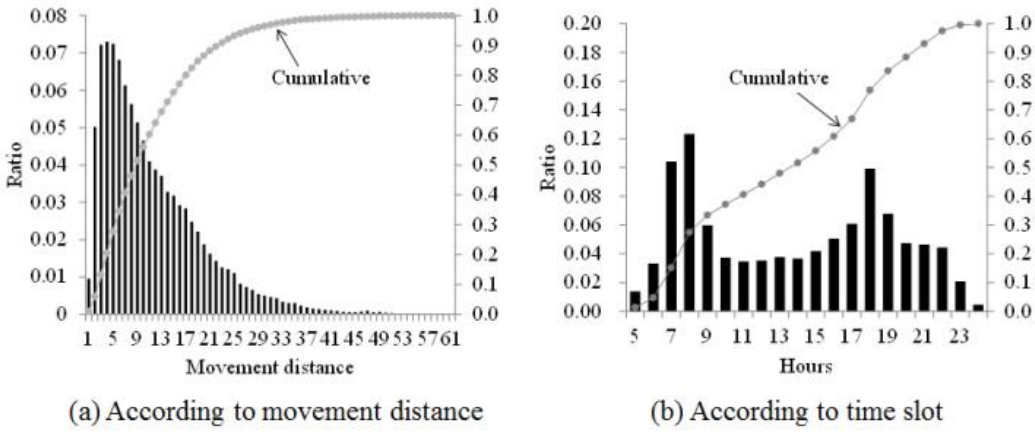

Figure 6. Observed movement distributions.

4.2. MZP identification

In this section, we present how MZPs are identified by merging MZPs from an OD dataset based on the proposed algorithm. Figure 7 illustrates the changes of the identified MZPs during the iteration steps of the proposed algorithm. Specifically, four graphs (a) to (d) in Figure 7 show the identified MZPs when the numbers of iteration steps were 0 (initial), 70, 140, and 193 (final), respectively. In each graph, the horizontal and the vertical axes are the origin and the destination stations, respectively. Since the total number of stations across the four lines was 148, each graph contains 21,904 cells each of which is associated to an origin and a destination stations. And, the colour of a cell shows the number of observed movements from its associated origin and destination stations. An MZP is shown as a rectangle because each MZP is represented as a set of horizontally and vertically adjacent cells.

As shown in Figure 7, through conducting MZP merges, the number of remaining MZPs was decreased while better MZPs in terms of coverage were obtained. The respective numbers of the remaining MZPs for the iteration steps, depicted in Figures 7 (a) to (d), were 2,391, 2,159, 1,911, and 1,745, respectively. On the other hand, at each iteration step, the best MZPs in terms of coverage were able to explain 12,540, 57,102, 71,262, and 73,926 observations, respectively. After finalizing iteration steps, most significant MZPs were discovered on the V-line or the O-line, as shown in Figure 7 (d). Since more observed movements on a line imply more opportunities to produce better MZPs in terms of coverage, the significantly large number of movements observed on the two lines compared to others might be contributed to yield such results. The best MZP in terms of coverage was {O39, O40, O41, O42} ⟶ {O31, O32, O33, O34, O35, O36, O37, O38}, shown in a box in Figure (d), whose coverage value was 1.552%.

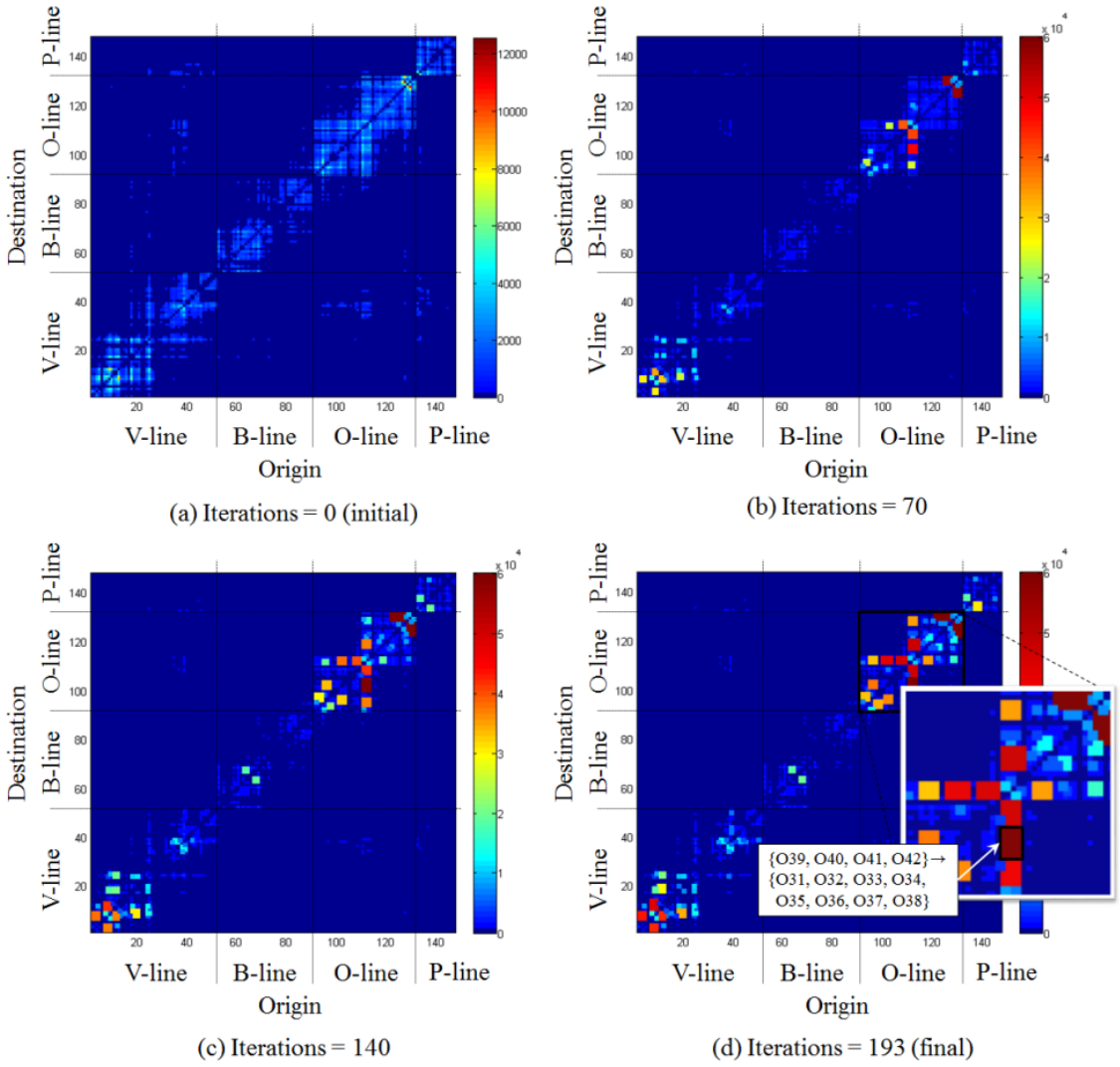

Figure 7. MZPs identified with the proposed algorithm according to iteration steps.

The MZPs identified from the collected OD dataset were then evaluated by using the three proposed measures, *v*-value, *a*-value, and *c*-value, and the top 10 MZPs in terms of *c*-value are presented in order in Table 2, and the results are visualized on a map of Seoul, shown in Figure 8.

The top 10 MZPs were able to cover 8.839% among the entire observed movements. The best MZP in terms of *c*-value, p_1, was capable to cover 1.552% of the entire observed movements, and its accuracy was 5.659. Even though its accuracy was lesser than those of some top ranked MZPs, it turned out that the MZP outperformed the others when both coverage and accuracy were considered. Interestingly, MZPs, p_8 and p_9, whose respective ranks in terms of coverage were 304 and 317 (their *v*-values were only 0.402 and 0.388, respectively), were finally ranked at the 8-th and 9-th positions

attributed to their high *a*-values, implying that the consideration of the trade-off between the amount of movements and the dependency between zones (i.e. coverage and accuracy) can be helpful to unveil serendipitous MZPs.

Additionally, we found some pairs of MZPs each of whose origin and destination zones were in opposite directions such as (p_3, p_4) and (p_8, p_9) in Table 2. For instance, the origin zone of p_3 is the same as the destination zone of p_4, and the destination zone of p_3 is same as the origin zone of p_4. These results might be yielded by the movement patterns which are mainly caused by commuting behaviours of people between zones. Moreover, there were pairs of MZPs such as (p_3, p_7) and (p_4, p_6) that share either origin or destination zones with each other. In particular, a zone that contains V6, V7, V8, and V9 was performing as an origin zone of different MZPs, p_3 and p_7, while the zone was also acting as a destination zone for different MZPs, p_4 and p_6, at the same time.

Table 2. Top 10 MZPs in terms of the combined measure (*c*-value).

MZP	Origin zone	Destination zone	# of covered movements	Measures		
				v-value (%)	*a*-value	*c*-value
p_1	O39, O40, O41, O42	O31, O32, O33, O34, O35, O36, O37, O38	73,926	**1.552**	5.659	**0.296**
p_2	O32, O33, O34, O35, O36, O37, O38	O39, O40, O41, O42	60,996	**1.280**	6.122	**0.280**
p_3	V6, V7, V8, V9	V1, V2, V3, V4	37,824	0.794	7.193	**0.239**
p_4	V1, V2, V3, V4	V6, V7, V8, V9	37,252	0.782	7.261	0.238
p_5	P1, P2, P3, P4	P5, P6, P7, P8	27,634	0.580	**7.934**	0.215
p_6	V10, V11, V12, V13, V14	V6, V7, V8, V9	44,755	0.939	4.847	0.213
p_7	V6, V7, V8, V9	V10, V11, V12	37,391	0.785	4.873	0.196
p_8	B11, B12, B13	B15, B16, B17	19,172	0.402	**8.626**	0.186
p_9	B15, B16, B17	B11, B12, B13	18,477	0.388	**8.803**	0.185
p_{10}	O9, O10, O11, O12, O13, O14, O15	O21, O22, O23, O24	63,629	**1.336**	2.469	0.182
sum	-	-	421,056	8.839	-	-

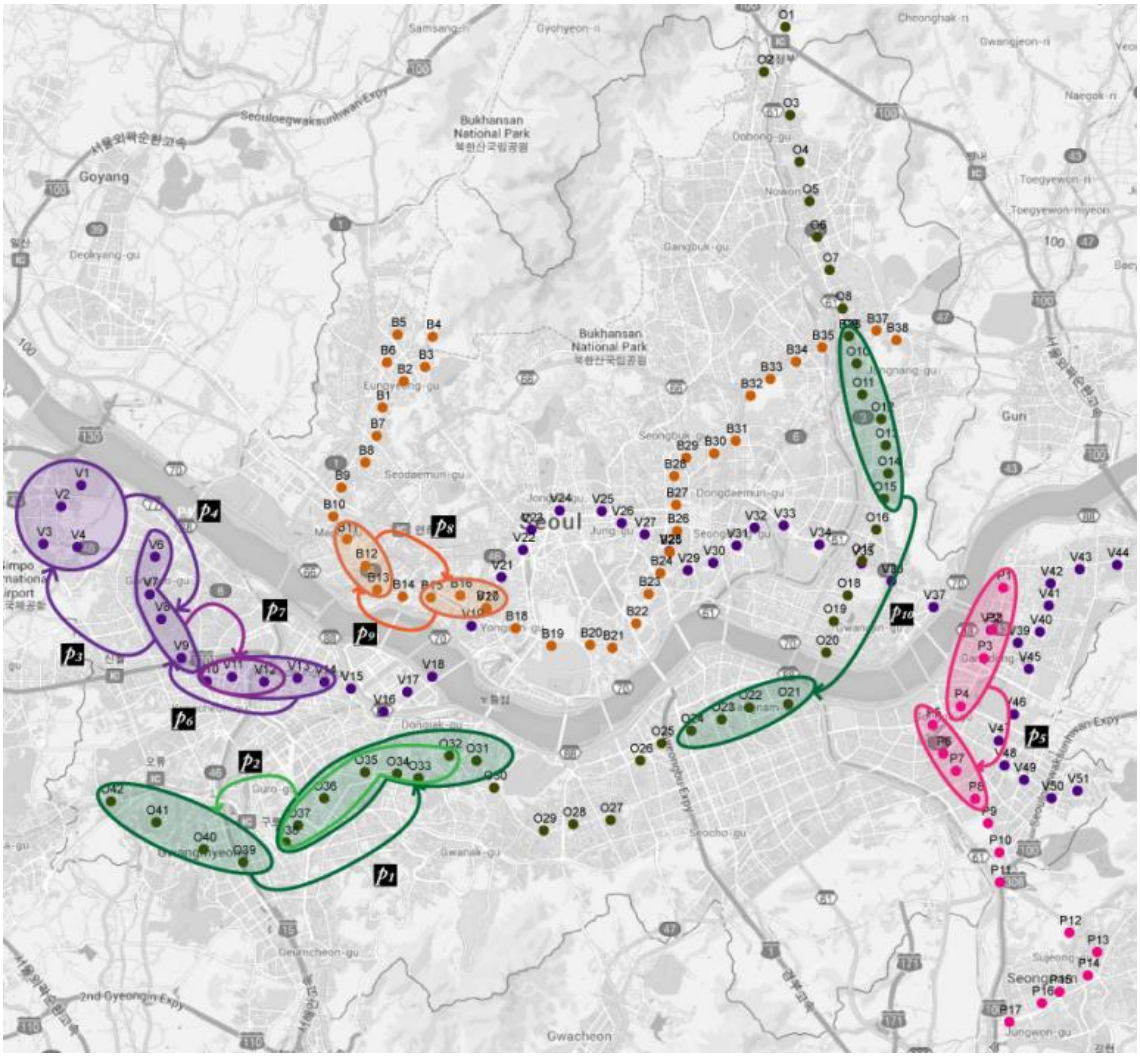

Figure 8. Visualization of the identified top MZPs.

4.3. MZP distributions

Figure 9 depicts the distribution of the distances between zones in the 1,745 discovered MZPs. The distances between zones in the discovered MZPs ranged from 1 to 49. The graph shows that more than half of the distances between zones associated to an MZP were less than 8 and the most frequently observed distance was 2. There hardly existed MZPs in which the distance between zones were larger than 20. This implies that closer zones are more likely to be related with each other. Here, the distance of an MZP was calculated as the average distance between stations in the origin zone and stations in the destination zone.

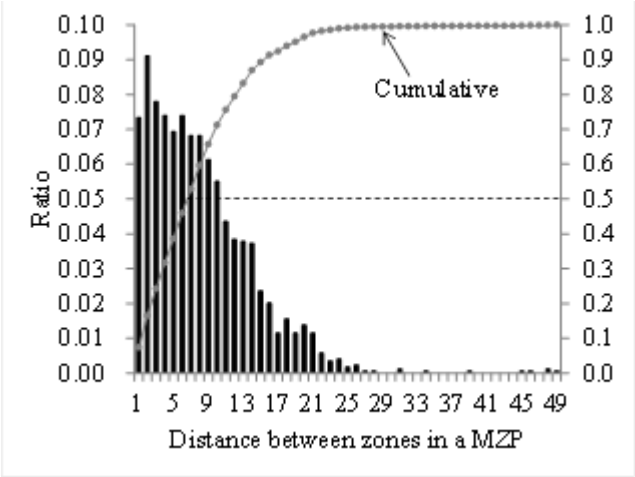

Figure 9. Distribution of the distances between zones.

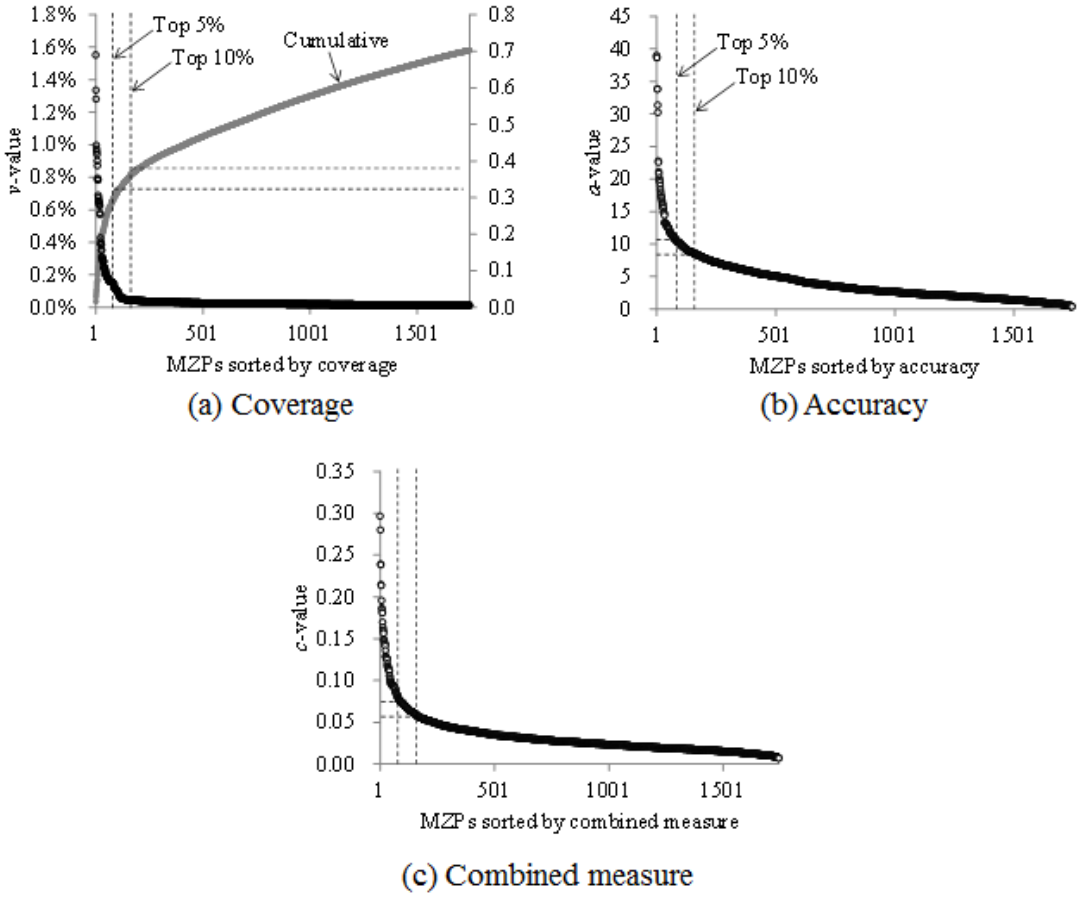

Figure 10. Distributions of the identified MZPs with respect to the suggested measures.

Figure 10 shows the distributions of the identified MZPs with respect to three suggested measures, indicating that only a small number of identified MZPs were significant. First, the *v*-values and the cumulative proportion for the MZPs are presented

in Figure 10 (a). The top 5% of the identified MZPs in terms of coverage, which included 87 MZPs, covered 32.566% of the entire observed movements while the top 10% of MZPs covered 38.059%. Next, Figure 10 (b) depicts the distribution of the *a*-values for the identified MZPs. Similar to coverage, the *a*-values for the top 5% MZPs were much higher compared to those of the others. The accuracy values for the 5-th and the 10-th percentiles of the identified MZPs were 10.394 and 8.348, respectively. Finally, Figure 10 (c) presents the distribution of the identified MZPs in terms of the combined measure. The *c*-values for the 5-th and the 10-th percentiles of the discovered MZPs were 0.077 and 0.068, respectively.

4.4. Commuting pattern analysis

We further investigated the top 10 MZPs, presented in Table 2, in terms of commuting behaviours of people. We assumed that the morning and the evening commutes can be captured mainly in the two rush hours that were mentioned in Section 4.1. That is, time windows of 7:00 to 10:00 and 17:00 to 20:00 were considered the morning and the evening commutes, respectively.

Figure 11 shows the comparison results between morning and evening commutes of the top 10 MZPs for three measures, *v*-value, *a*-value and *c*-value. Some MZPs such as p_{10} and p_1 outstand in morning commute, compared to evening commute. For instance, MZPs p_{10} and p_1 have much higher coverages (i.e. *v*-values) of morning commute than those of evening commute, as shown in Figure 11 (a), although both accuracies in their morning and their evening commutes were almost the same, as shown in Figure 11 (b).

On the other hand, some other MZPs were more suitable to address movements for evening commute than morning commute. MZPs p_2, p_6 and p_4 have much higher coverages of evening commute than those of morning commute, as shown in Figure 11 (a), although their accuracies in the morning a little higher than those in the evening commute. For the rest of MZPs, there was no significant tendency according to commuting behaviours, implying that they were playing for both morning and evening commutes.

Interestingly, the accuracy values of all the identified top MZPs for morning commute were higher than those for evening commute. This indicates that the commuting behaviours of people in the morning are more patternized compared to those in the evening.

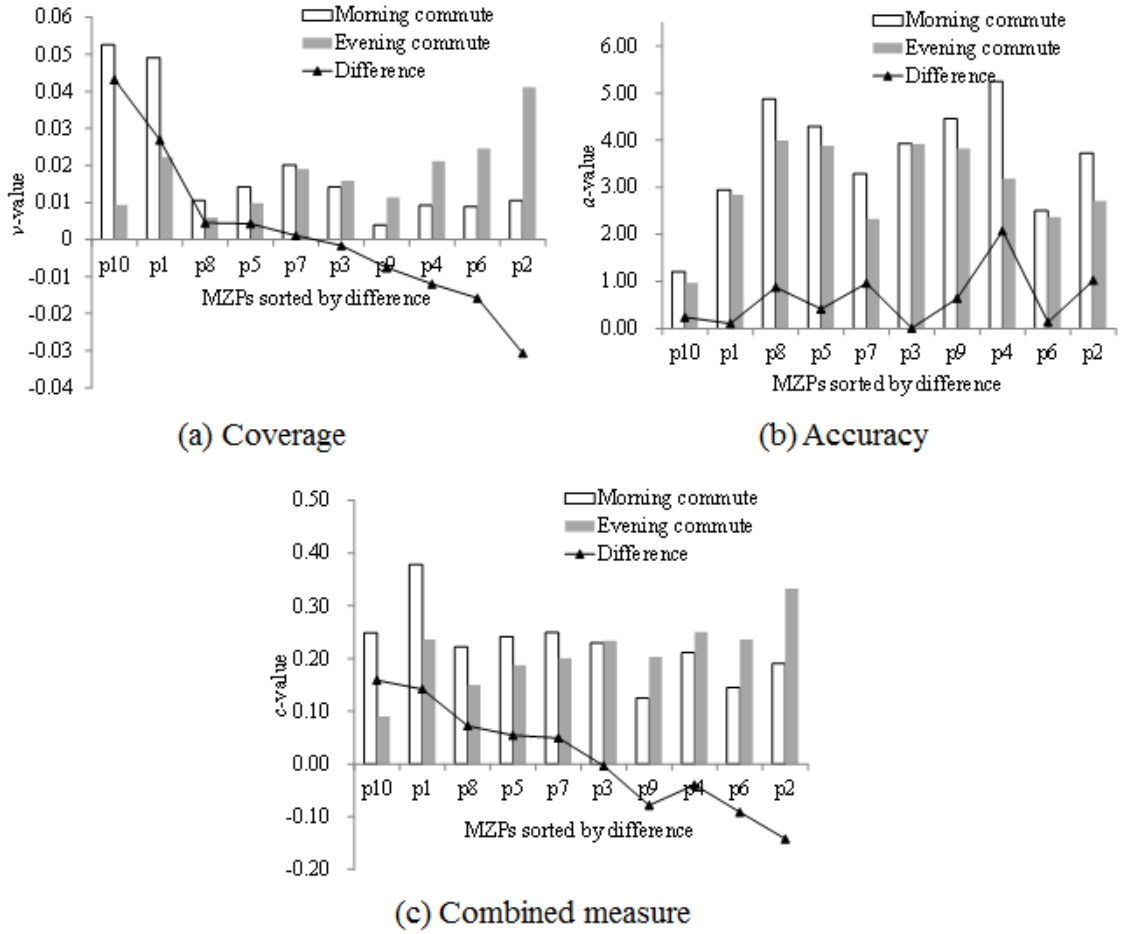

Figure 11. Comparison of the discovered top MZPs according to morning and evening commuting behaviours.

5. Discussion and conclusions

In this research, we suggest a data driven approach to identify movement patterns of people between zones, called MZPs, based on boarding behaviours of people in subway networks. Specifically, our approach attempts to simultaneously identify zones and movement patterns between zones from an OD dataset unlike the previous approaches that separately investigate the two issues. We adopted the proposed approach in a real-world OD dataset obtained from an EPS in a subway network in Seoul, Korea. Throughout the experiments, our approach showed satisfactory results in identifying strong MZPs for supporting practitioners to better understand the existence of zones and the relations between them.

Our approach included both identification and measurement of movement patterns based on generic information such as origins, destinations, and adjacent stations. It is believed that the proposed method can also be applied in other types of transportation networks such as bus and taxi. Moreover, discovered MZPs are able to facilitate more precise transportation planning and route reorganization by capturing the actual dependencies among regions in terms of people's movements. For instance, non-stop bus routes and other alternative transportation means between zones associated with a MZP can be developed. We also plan to conduct research on more precise evaluation of MZPs by additionally considering demographic features.

References

Antikainen, J., 2005. The concept of functional urban area. *Findings of the Espon Projects*, 1 (1).

Bagchi, M. and White, P.R., 2004. What role for smart-card data from bus systems?. *Municipal Engineer*, 157 (1), 29-46.

Bagchi, M. and White, P.R., 2005. The potential of public transport smart card. *Transport Policy*, 12 (5), 464-474.

Bhat, C., 2001. Modeling the commute activity-travel pattern of workers: formulation and empirical analysis. *Transportation Science*, 35 (1), 61-79.

Blythe, P., 2004. Improving public transport ticketing through smart cards. In *Proceedings of the Institute of Civil Engineers, Municipal Engineer*, 47-54.

Chu, K.K.A. and Chapleau, R., 2008. Enriching archived smart card transaction data for transit demand modeling. *Journal of the Transportation Research Board*, 2063 (1), 63-72.

Chu, K.K.A. and Chapleau, R., 2010. Augmenting transit trip characterization and travel behavior comprehension. *Journal of the Transportation Research Board*, 2183 (1), 29–40.

Fusco, G. and Caglioni, M., 2011. Hierarchical Clustering through spatial interaction data. The case of commuting flows in south-eastern France, *Lecture Notes in Computer Science*, 6782, 135-151.

Ghasemzadeh, M., Fung, B., Chen, R. and Awasthi, A., 2014. Anonymizing trajectory data for passenger flow analysis. *Transportation Research Part C*, 39, 63-79.

He, B., Ding, Y. and Yan, E., 2012. Mining patterns of author orders in scientific publications. *Journal of Informetrics*, 6 (3), 359-367.

Hoffman, M., Wilson, S.P. and White, P., 2009. Automated Identification of linked trips at trip level using electronic fare collection data. In *The Transportation Research Board 88th Annual Meeting*, 09-2417, Singapore, 840-845.

Jang, W., 2010. Travel time and transfer analysis using transit smart card data. *Journal of the Transportation Research Board*, 2144 (1), 142-149.

Joh, C.-H., Arentze, T.A. and Timmermans, H. J.P., 2001. Multidimensional sequence alignment methods for activity-travel pattern analysis: A comparison of dynamic programming and genetic algorithms. *Geographical Analysis*, 33, 247–270.

Karlsson, C., 2007. Clusters, functional regions and cluster policies. *JIBS and CESIS Electronic Working Paper Series*, 84.

Konjar, M., Lisec, A. and Drobne, S., 2010. Method for delineation of functional regions using data on commuters. In *Proceedings of the 13-th AGILE International Conference on Geographic Information Science*, Guimarães, Portugal, 10.

Lee, K. and Park, J.S., 2005. Traversal pattern analysis of transit users in the metropolitan Seoul. In *Proceedings of International Forum on the Public Transportation Reform in Seoul*, Seoul.

Lee, S. and Mark D, H., 2011. Travel pattern analysis using smart card data of regular users. In *Transportation Research Board 90-th Annual Meeting*, 11-4258.

Liu, L., Hou, A., Biderman, A., Ratti, C. and Chen, J., 2009. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In *the 12-th International IEEE Conference on Intelligent Transportation Systems*, 1-6.

Ma, X., Wu, Y. J., Wang, Y., Chen, F. and Liu, J., 2013. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C*, 36, 1-12.

Martin, D., 2003. Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, 17 (2), 181-196.

Morency, C., Trépanier, M. and Agard, B., 2006. Analysing the variability of transit users behaviour with smart card data. In *the 9-th International IEEE Conference on Intelligent Transportation Systems Conference (ITSC06)*, 44-49.

Moreno-Regidor, P., García López de Lacalle, J. and Manso-Callejo, M., 2012. Zone design of specific sizes using adaptive additively weighted Voronoi diagrams. *International Journal of Geographical Information Science*, 26 (10), 1811-1829.

Munizaga, M., Palma, C. and Mora, P., 2010. Public transport OD matrix estimation from smart card-data. *Transportation Policy*, 14 (3), 193-203.

Munizaga, M. A. and Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C*, 24, 9-18.

Park, J.Y., Kim, D.J. and Lim, Y., 2008. Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea, *Journal of the Transportation Research Board*, 2063 (1), 3-9.

Pelletier, M., Trépaniera, M. and Morency, C., 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C*, 19 (4), 557-568.

Srinivasan, S. and Ferreira, J., 2003. Travel behavior at the house level: understanding linkages with residential choice. *Transportation Research Part D*, 7 (3), 225-242.

Tan, P., Kumar, V. and Srivastava, J., 2004. Selecting the right objective measure for association analysis. *Information Systems*, 29 (4), 293-313.

Trépanier, M., Morency, C. and Agard, B., 2009. Calculation of transit performance measures using smartcard data. *Journal of Public Transportation*, 12 (1), 79-96.

Trépanier, M. and Morency, C., 2010. Assessing transit loyalty with smart card data. In *The 12-th World Conference on Transportation Research*, Lisbon, Portugal.

Yuan, J., Zheng, Y. and Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs, In *Proceedings of the18th ACM SIGKDD International Conference on Discovery and Data Mining*, ACM, 186-194.

Zhao, J., Frumin, M., Wilson, N. H. and Zhao, Z., 2013. Unified estimator for excess journey time under heterogeneous passenger incidence behavior using smartcard data. *Transportation Research Part C*, 34, 70-88.

Appendix A.

See Table 3.

Table 3. Full names of stations according to lines (Sym refers to symbol).

(a) V-line

Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name
V1	Banghwa	V2	Gaehwasan	V3	Gimpo Itn'l Airp.	V4	Songjeong	V5	Magok
V6	Balsan	V7	Ujangsan	V8	Hwagok	V9	Kkachisan	V10	Sinjeong
V11	Mok-dong	V12	Omokgyo	V13	Yangpyeong	V14	Y.D.P.-gu Office	V15	Y.D.P. Market
V16	Singil	V17	Yeouido	V18	Yeouinaru	V19	Mapo	V20	Gongdeok
V21	Aeogae	V22	Chungjeongno	V23	Seodaemun	V24	Gwanghwamun	V25	Jongno 3-ga
V26	Euljiro 4-ga	V27	D.D.M. H. Park	V28	Cheonggu	V29	Singeumho	V30	Haengdang
V31	Wangsimni	V32	Majang	V33	Dapsimni	V34	Janghanpyeong	V35	Gunja
V36	Achasan	V37	Gwangnaru	V38	Cheonho	V39	Gangdong	V40	Gil-dong
V41	Gubeundari	V42	Myeongil	V43	Godeok	V44	Sangil-dong	V45	Dunchon-dong
V46	Olympic Park	V47	Bangi	V48	Ogeum	V49	Gaerong	V50	Geoyeo
V51	Macheon								

(b) B-line

Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name
B1	Eungam	B2	Yeokchon	B3	Bulgwang	B4	Dokbawi	B5	Yeonsinnae
B6	Gusan	B7	Saejeol	B8	Jeungsan	B9	Digital Media City	B10	WorldCupStadium
B11	Mapo-gu Office	B12	Mangwon	B13	Hapjeong	B14	Sangsu	B15	Gwangheungchang
B16	Daeheung	B17	Gongdeok	B18	Hyochang Park	B19	Samgakji	B20	Noksapyeong
B21	Itaewon	B22	Hangangjin	B23	Beotigogae	B24	Yaksu	B25	Cheonggu
B26	Sindang	B27	Dongmyo	B28	Changsin	B29	Bomun	B30	Anam
B31	Korea Univ.	B32	Wolgok	B33	Sangwolgok	B34	Dolgoji	B35	Seokgye
B36	Taereung	B37	Hwarangdae	B38	Bonghwasan				

(c) O-line

Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name
O1	Jangam	O2	Dobongsan	O3	Suraksan	O4	Madeul	O5	Nowon
O6	Junggye	O7	Hagye	O8	Gongneung	O9	Taereung	O10	Meokgol
O11	Junghwa	O12	Sangbong	O13	Myeonmok	O14	Sagajeong	O15	Yongmasan
O16	Junggok	O17	Gunja	O18	Ch.Grand Park	O19	Konkuk Univ.	O20	Ttukseom Resort
O21	Cheongdam	O22	G.N.-gu Office	O23	Hak-dong	O24	Nonhyeon	O25	Banpo
O26	Express Bus Term.	O27	Naebang	O28	Isu	O29	Namseong	O30	Soongsil Univ.
O31	Sangdo	O32	Jangseungbaegi	O33	Sindaebangsamgeori	O34	Boramae	O35	Sinpung
O36	Daerim	O37	Namguro	O38	Gasan Dig. Comp.	O39	Cheolsan	O40	Gwangmyeongsageori
O41	Cheonwang								

(d) P-line

Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name	Sym	Full name
P1	Amsa	P2	Cheonho	P3	Gangdong-gu Office	P4	Mongchontoseong	P5	Jamsil
P6	Seokchon	P7	Songpa	P8	Garak Market	P9	Munjeong	P10	Jangji
P11	Bokjeong	P12	Sanseong	P13	Namhansanseong	P14	Dandaeogeori	P15	Sinheung
P16	Sujin	P17	Moran						