

Article

Flow Orientation Analysis for Major Activity Regions Based on Smart Card Transit Data

Parul Singh, Kyuhyup Oh and Jae-Yoon Jung *

Department of Industrial and Management Systems Engineering, Kyung Hee University, Yongin, Gyeonggi 17104, Republic of Korea; parul.singh2712@khu.ac.kr (P.S.); k8383@khu.ac.kr (K.O.)

* Correspondence: jyjung@khu.ac.kr; Tel.: +82-31-201-2537

Received: 30 July 2017; Accepted: 16 October 2017; Published: 21 October 2017

Abstract: Analyzing public movement in transportation networks in a city is significant in understanding the life of citizen and making improved city plans for the future. This study focuses on investigating the flow orientation of major activity regions based on smart card transit data. The flow orientation based on the real movements such as transit data can provide the easiest way of understanding public movement in the complicated transportation networks. First, high inflow regions (HIRs) are identified from transit data for morning and evening peak hours. The morning and evening HIRs are used to represent major activity regions for major daytime activities and residential areas, respectively. Second, the directional orientation of flow is then derived through the directional inflow vectors of the HIRs to show the bias in directional orientation and compare flow orientation among major activity regions. Finally, clustering analysis for HIRs is applied to capture the main patterns of flow orientations in the city and visualize the patterns on the map. The proposed methodology was illustrated with smart card transit data of bus and subway transportation networks in Seoul, Korea. Some remarkable patterns in the distribution of movements and orientations were found inside the city. The proposed methodology is useful since it unfolds the complexity and makes it easy to understand the main movement patterns in terms of flow orientation.

Keywords: smart card transit data; flow orientation analysis; public transportation network

1. Introduction

Most major cities in the world have gradually developed over time without pre-defined plans. This often causes unarranged and unexpected changes of many regions in the city. Diagnosis of the urban sprawl is crucial to understand public movements, which have many facets attached: primarily traffic congestion, locating business centers, and type of residential setup; and secondarily information flow, spread of biological viruses, and urban and transit planning [1–7]. The objective of this paper was to develop a methodology for understanding the flow orientation of public movements from real data sources such as smart card transit data. Even though many methods for public movement analysis have been developed, one of the easiest ways of understanding the movement flow is to show the directional flow of each region. Moreover, the flow orientation analysis based on the smart card transit data, which contain the exact movement information of the public in transportation networks according to time and location, has not been analyzed yet. Visualization of such directional movements on the map is very useful for easy understanding of real public movements which were distributed in complicated transportation networks in the developed cities.

Generally, major activities in a city are concentrated in a small number of regions. By interpreting the flow towards the major activity regions, one can understand most of the public movements in the city. In this study, the term *High Inflow Regions* (HIRs) is used to refer the regions that attract the majority of the public to perform their activities during the daytime or residential areas at night.

The term is similar to the poly-centers used in other studies [2,8], but HIRs also contain the residential regions, as well as major areas for daytime activities. Meanwhile, high traffic is plausible in network routes that connect origins and destinations; origins and destinations of trips along with time of travel depend on living and working locations. The HIRs for major daytime activities can be determined by investigating trips for morning commutes, while the HIRs for residences can be derived from the trips for evening commutes. The two kinds of HIRs are called morning HIRs and evening HIRs in this study, respectively. These are significant in urban development projects, transportation improvement, and social network analysis.

Another aspect covered in this paper was the distribution of traffic from the origin regions towards discovered HIRs. This flow orientation in traffic was measured from each direction towards HIRs. This directional flow orientation provided the comparative incoming flow from regions in different directions. Transport service providers can benefit by concentrating on dominant directions to provide the optimum level of transport for those regions. The availability of public transport can also be inspected for the least inflow directions.

In this study, the flow orientation was intended to be analyzed using real transit data of public transport. Smart card transit data was very useful to investigate the detailed movements in traffic according to specific times and locations. More specifically, the transit data in the morning and evening peak hours were focused on in order to identify the morning and evening HIRs in a city, respectively. For the purpose of abstraction and simplicity of regions, the Geohash system was adopted, which is a geo-coding system of mapping a specific location in the world to a unique code according to the required resolution [9].

A directional inflow vector was obtained for each HIR based on the direction of the trip to the destination, and the vector was used to compare the similarity between two HIRs. Furthermore, the flow orientation patterns of HIRs were visualized by clustering the HIRs based on their directional inflow vectors. To illustrate the proposed methodology, smart card transit data for the bus and subway networks in Seoul, Korea, were utilized, and the implication of the method was described with the experimental results. In the experiments, the flow orientation patterns of morning HIRs were more variable compared to those of evening HIRs, which meant the working regions were more concentrated than the living regions in terms of orientation in this city. It was also possible to find the relationship between flow orientation and regional environment, such as a river and expressway in the city.

The contributions of this study can be summarized as follows:

- To understand major activity regions in a city, morning and evening HIRs were derived and investigated from real transit data.
- To provide a comprehensible way of showing public movement, the method for flow orientation analysis were developed with a directional inflow vector and a dominant factor.
- To lessen the complexity of complicated transportation networks, the Geohash system was adopted for scalable abstraction of bus stops and subway stations.
- To show similar flow orientation patterns of HIRs, hierarchical clustering of HIRs was applied and then visualized on a city map.
- Through a smart card transit data in Seoul, the illustration and effectiveness of the proposed method for flow orientation analysis were presented.

The remainder of this paper is structured as follows. In Section 2, the related work is discussed. In Section 3, the methodology for processing smart card transit data, discovering HIRs, and analyzing flow orientation of these regions is presented. In Section 4, the proposed methodology for flow orientation analysis is demonstrated with a smart card transit data collected in Seoul, Korea. Section 5 makes concluding remarks and describes future work.

2. Related Work

The subject of our research is use of public transport data such as smart card transit data to study human mobility for orientation of flow. Hence, a literature review on smart card-based transit data

analysis will be provided and the studies will be segregated similar to our research interest to help readers understand and compare the research purposes of studies based on the smart card transit data. After that, the flow analysis on human mobility will be reviewed and similar studies to our method will be introduced, although their data sources were not smart card transit data because of little work on flow orientation based on the smart card transit data.

A review paper on smart card data-based study [10] focused on smart card technology applied in public transit as a whole, emphasizing research classification. This paper helps us develop perspective in research direction. We found few related studies conducted in past on smart card data and we listed them in Table 1. Since the travel behavior using human mobility was studied for various purposes by researchers, we found it well again to categorize them for the purpose of the studies. Four categories of the studies were presented in Table 1 and described in succeeding paragraphs.

The first category is the clustering of geographical areas, which was also used in our research to group major activity regions for flow orientation. Spatial and temporal regularity of travelers was measured by researchers in the past by grouping them by chosen boarding/alighting stops and routes on different weekdays, and by grouping them by time of travel [11–13]. Morency et al. were further interested in the class wise regularity patterns of travelers [11,12]. Kieu et al. went on to categorize passengers as based on the regular selection of time of travel over observed time frame [14]. The categories are regular, habitual and irregular passengers occasional. Some researchers have considered the travel behavior for a geographical location. Kim et al. clustered subway stops and created zones having similar directions of travel [15]. Du et al. clustered regions and studied travel patterns between regions regarding direction and destination of travel [16]. These existing researches have adjacency as common constraint that affects the accuracy of derived results. Another similar study was conducted by Roth et al. [5]. They studied variation in flow between subway stops and the orientation of flow. Our work is based on their study. We took it a step further by developing a more effective technique for studying the concentration of flow to perform flow orientation analysis in a detailed manner.

Table 1. Related studies on human mobility using smart card transit data.

Category	Subject	Method	Transit Data
Geographical clustering	Travelers for age occupation wise travel behavior [11]	k-means	277 consecutive days
	Travelers for regularity in boarding [12]	k-means	277 consecutive days
	Mining travel patterns [13]	DBSCAN	5 consecutive weekdays
	Origin-destination pairs for discovering zones based on movement patterns [15]	Clustering	5 consecutive weekdays
	Travelers for direction and destination of travel [16]	DBSCAN	92 days
	Subway stations for high inflow poly-center [4]	k-means	1 week
	Travelers for temporal boarding pattern [14]	DBSCAN	4 months
Movement visualization	Drawing travel trajectory and visual representation of movement pattern [17,18]	co-map	5 days
	Job housing location and commuting pattern [19]	GIS platform	7 days
	Interactive visualization of human mobility with activity context [20]	-	1 week
Relationship extraction	Relationship between mobility pattern of individual and daily activities [21]	Bayesian classifier	1 week
	Travel behavior analysis by measuring passenger turnover [22]	rule-based method	Approx. 3 years
	Behavioral trip purpose estimation [23]	Bayesian classifier	20 months
	Relation of arrival and departure at certain station [20]	-	1 week
Flow orientation	Discovering flow orientation for poly-centers [5]	compass direction	1 week
	Discovering spark regions based on high density routes [16]	DBSCAN	92 days
	Identifying activity centers and clustering them for spatial proximity and temporal flows [24]	clustering	1 day
	Identifying industrial agglomerations and their orientation with respect to different modes of transport to check importance of transport accessibility [25]	-	1 day

Researchers interested in studying travel behavior generally use visualization for studying movement patterns. Prior to this study, maps were used to visualize commuting patterns at a time of day for different categories of passengers [17,18], and job and housing locations were marked in the

form of stations or bus stops [19]. Map visualization also evolved over a period of time with more granularity in visualization combined with visual analytics. Zeng et al. used a map visualization with multidimensional attributes such as the volume of arrivals and the departure for each day of a week and activity categories in the area [20]. Compared to the previous studies on movement visualization, we showed HIRs and similarity between them in term of orientation pattern on the map.

Another research trend is to derive relationships between human mobility and the reason of travel from smart card transit data. Existing studies that fall into this category either relate passenger journeys with their travel purpose [21] using additional survey data; or relate the attractiveness of bus stops, stations or travel modes in terms of passenger inflow [22,23]. Zeng et al. linked the arrivals and departures at stations with activities in the area based on the time of arrivals and departures [20]. Following this trend, our efforts were first to extract the relationship between regular traveler movements with the concentration of day and night activities; and second to extract the relationship between two regions for similarity in orientation of inflow.

The analysis of flow orientation, which was the objective of our research, was not very popular in our literature review of smartcard data analysis. Du et al. touched on this topic and analyzed routes connecting high density regions [16]. Roth et al. grouped bus stops based on proximity and high inflow in their research, and analyzed the orientation of flow for them [4]. Cats et al. identified public transport activity centers in line with passenger flow and spatial proximity [24]. They studied total flow (the sum of arrivals and departures), differential flow, and flow ratio variation in these clusters for a time of day segments. In another study, Song et al. identified the type of industrial agglomerations and analyzed each of these orientations with a respect to different transportation modes assess relation with transport accessibility [26]. This study gauged the proximity of agglomerations to different transport modes to access the source of flow and importance of transport accessibility for industry types. They studied the orientation of flow with the variation that we grouped the discovered regions for flow orientation. Furthermore, the implementation of their methodology in other geographical areas was less feasible. We performed analyses by defining new measures for comparing flow orientation, which was useful for comparison between two flow orientations. Moreover, we also demonstrated the use of clustering and visualizing tools in analyzing flow orientation geographically.

Furthermore, while most previous studies analyzed the smart card transit data at the bus stop/station or route level [2], our research concentrated on the data at the geographical level. One of the highlights of this research is to show that the geographical complex flow information was analyzed in a simple way. The motivation for this was the assumption that analyzing the flow of high inflow areas is ample enough to understand rough geographical movement, and also to plan for future urban and network development. Hence, this study focused on the orientation of incoming flow for just a small number of HIRs. Moreover, the Geohash system was adopted to abstract the information of the geographical locations of stations and bus stops, and to choose a proper resolution. Using the Geohash system, we reduced the complexity of flow orientation analysis in this study.

The scope of this paper was restricted to analyze the flow orientation based on smart card transit data. Due to little work on past smart card data-based analysis, some major studies on human mobility from other types of data sources such as mobile phones and car GPS data were briefly introduced to compare with our study. To provide a glance on urban human mobility, which is an emerging area of interest for researchers, a few examples are provided. Zhu et al. clustered trips based on similarities of origins and destinations to bundle flow. This study provided a flow mapping view from different levels of resolutions and demonstrated the method with taxi data [25]. Wu et al. studied the urban human mobility by grouping trips with the same destinations [27]. They called it co-occurrence and the approach helped to understand the traffic flow at a given time in a given area. In their study, mobile phone trajectory data was used for demonstration. Andrienko et al. segmented events from trajectories based on car GPS data using specific query instead of focusing on the whole travel trajectories [28]. These clusters of movement events were used to analyze traffic conditions in the areas of interest. Those studies tried to analyze the traffic flow and visualize them geographically.

In a similar way, this study also analyzed flow patterns in terms of flow orientation by using smart card transit data to discover the movement patterns of the public.

Meanwhile, there are three studies that were most related to this study [5,15,16]. Each of these studies attempted to develop methods to apply smart card transit data for analyzing high density concentration of flow and public mobility patterns. It seems that not many researchers have focused on performing this task. The HIR clusters in this study were a similar concept to the spark regions in [16], poly-centers in study [5], and MZPs in [15]. Each of the three studies discussed peak hour flow. However, they did not use it exclusively for discovering the high density of flow. Considering all of the day data simultaneously for discovering flow concentration could be misguided, since it includes both the journey and return. To remove this bias, we analyzed peak hour data, which served the purpose of analyzing regular travel behavior and could better help in capacity evaluation.

Morning and evening peak commute data were analyzed using the same method in this study. The former was for work destinations and the latter for residential. Poly-centers described in [5] did not separate residential and working destinations based merely on data. Spark regions in [16] could be bifurcated as high inflow or outflow, and considered both inflow and outflow for the entire day, but they could not clearly separate out residential and work places. In their experiments in study [15], the MZPs were discovered for the morning and evening commute, while the MZPs could not be distinguished for residential or work concentration areas.

The normalized directional flow vector along with a dominance factor that was introduced in our study can help represent the variation in contributed flow for HIRs. It can also be used to compare any two HIRs for dominant direction and balance in the flow orientation. Roth et al. [4] performed directional flow analysis in which the normalization was through a null model, having actual inflow and outflow degrees of stations and randomized rides. Their method was not proper for comparing the characteristics of two regions since it targeted flow orientation for a required region.

3. Methodology

Smart card transit systems provide the ability to examine user behavior in a better way than revealed preference surveys [29]. The usage of transit systems in changing urban movement and local communities can be monitored in more precise and flexible manners [5,30]. Taking this into consideration, smart card data was used for spatial analysis of a metropolitan city.

This smart card transit data were processed to transform riding and alighting stations/stops to the required resolution area using Geohash codes. Then, actual trips were segregated from transfer to obtain the final origin and destination (OD) database. This OD database was analyzed in this study to discover HIRs, which were the spatial regions on the map where the majority of the population preferred to travel. Each of these HIRs were further analyzed for flow orientation in this study considering inflow from eight compass directions. The analysis was carried out by comparing the directional flow for balance in the flow orientation.

3.1. Data Pre-Processing

The dataset considered for analysis in this study consisted of smart card transit data. It contained OD data in the form of station codes with time stamps. Other various attributes, such as amount paid, mode of transport, passenger category, route, and direction, were present in the smart card dataset.

During pre-processing, the origin and destination in the form of station codes were transformed to Geohash codes using GPS location. The main attributes of the OD dataset used in this study are presented in Table 2. The origin and destination bus stops or subway stations were mapped to Geohash codes using their longitude and latitude. Although the Geohash system was not requisite in this research, it dramatically reduced the complexity of dealing with a large number of subway stations and bus stops. Geohash can globally offer the possibility to increase or decrease the size of considered regions by adjusting the number of digits according to the required analysis precision. Geohash codes range from 1 to 12 digits according to the different precision. As an example, Figure 1 depicts how

a 4-digit grid, wydm, representing approximately $39.1 \text{ km} \times 19.5 \text{ km}$ area of Seoul, is divided into 32 smaller 5-digit grids, and a 5-digit grid, wydmb, is again divided into 32 6-digit grids. In this study, the transportation data used for the analysis were processed to 6-digit Geohash codes at a spatial resolution level. In addition, we ignored transfers to reflect only the final destinations. To achieve this, trips where the stoppage between two trips was less than 30 min were merged, and the first origin and the final destination were then used.

Table 2. Basic attributes of smart card transit data.

Attribute	Description	Data Type
Passenger code	Smart card serial number	Numeric
Origin station	Origin station number (card punch in)	Numeric
Boarding time	Boarding date and time at origin station	Date time
Destination station	Destination station number (card punch out)	Numeric
Alighting time	Alighting date and time at destination station	Date time

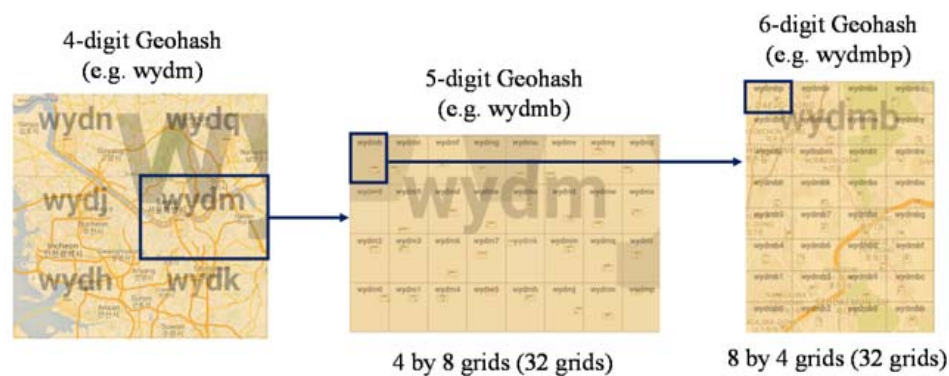


Figure 1. Hierarchical structure of the Geohash system for global geo-coding.

3.2. Discovery of High Inflow Regions

The pre-processed data contained origin and destination regions with time stamps for each individual trip. The destination regions were analyzed to find the HIRs, which were typically preferred destinations for the majority of travelers. The high inflow destination regions were chosen using the Pareto principle [31,32]. The Pareto principle, which is also known as the 80/20 rule, acted as a basis for segregating HIRs. Typically, around 20 to 30% of regions that contributed to 80% of the total inflow were selected as HIRs and further analyzed for flow orientation. The proposed rule, also called the law of the vital few, was justifiable for assigning HIRs, since it is widely used in various businesses and scientific fields for making decisions. The assumption was that analyzing those around 20% alighting regions could explain the flow orientation of most of the area by overlooking noise.

3.3. Flow Orientation Analysis

Flow orientation in this study signified the directional ratio of the movement amounts of passengers from origins to the specific destination. To investigate the flow orientation of a HIR (destination), all origin regions of the destination were divided with respect to relative direction to the target destination. In this study, an arrow from the center of the destination to the center of the origin was drawn virtually and then the angle of the arrow was used to map the direction of the origin to one of eight compass directions.

The directions of origin regions of a HIR and the flow amounts toward the HIR were used to measure the directional contribution of flow for the HIR. The directional contribution of flow for a particular HIR was proportional to the total incoming flow from all of the origin regions that belonged to the corresponding direction. Specifically, the directional inflow was used to measure each directional

contribution, and the directional inflow of HIR_i from direction d , denoted by f_i^d , was calculated as Equation (1):

$$f_i^d = \sum_{R_j \in O^d(HIR_i)} f_{ji} \quad (1)$$

$$f_i^{max} = \max_d f_i^d \quad (2)$$

where $O^d(HIR_i)$ is a set of regions R_{j_s} that are located in the d -th direction of HIR_i , and f_{ji} is the movement amount from R_j to HIR_i . In Equation (2), f_i^{max} is the maximum inflow of HIR_i among all of the directional inflows of HIR_i .

To compare the flow orientations among HIRs, the normalized directional inflow vector of HIR_i , denoted by F_i , was calculated. The vector was derived from all of the directional inflows of HIR_i normalized by the maximum directional inflow, as shown in Equation (3).

$$F_i = \left(\frac{f_i^1}{f_i^{max}}, \dots, \frac{f_i^D}{f_i^{max}} \right) \quad (3)$$

The normalized vector could be applied to measure the similarity among HIRs in terms of flow orientation. In this equation, D is the number of considered directions and eight compass directions (i.e., $D = 8$) were used in this study.

After the normalized directional inflow was derived for each HIR, we also measured the imbalance in inflow contributed by each direction. The dominance factor df was introduced to evaluate the dominance of the maximum directional inflow for other inflows as follows:

$$df_i = \frac{\sum_{d=1}^D (f_i^{max} - f_i^d)}{(D-1) f_i^{max}}. \quad (4)$$

In Equation (4), the dominance factor of HIR_i , denoted by df_i , is the summation of the differences between the maximum inflow and all of the directional inflows divided by the $D-1$ times of the maximum directional inflow. In flow orientation analysis, the dominance factor in terms of direction measured the extent to which the orientation of inflow was dominated by a single direction.

df had the maximum value of 1 when all of the directional flows were zero except for the maximum direction, while it had the minimum value of 0 when all of the directions equally contributed to the inflows. In other words, a high value of df indicated that a few directions contributed towards the inflow more than the other directions, while a low df indicated that all of the directions contributed to the inflow similarly.

4. Experiments

The flow orientation analysis was illustrated by applying the proposed methodology to smart card transit data in Seoul, Korea. In Seoul, the public transportation means such as bus and subway have charge of 64.3% among all transportation means, and the adoption rate of the smart card for public transportation reaches around 99%. A single weekday was chosen to analyze the regular travel patterns since it was known that weekdays had similar movement patterns in our preliminary study. The overall procedure of data transformation and experiments are presented in Figure 2.

The bus and subway transit data in Seoul on a chosen weekday were used as input for the analysis. The data was first processed by converting stop/station numbers to Geohash codes and distinguishing transfers from trips. Later, we extracted the transit data in the morning and evening peak hours from the transformed data having the origin and destination in the form of Geohash codes, which was proper for analyzing major movement patterns in a city. This transformed data of morning and evening peak commute hours was then used to discover the HIRs where most of the activities were concentrated during the day and night. Finally, the HIRs were clustered using agglomerative

hierarchical clustering (AHC) to investigate similar flow orientation patterns of HIRs. Each experiment and the results are presented in the following sub-sections.

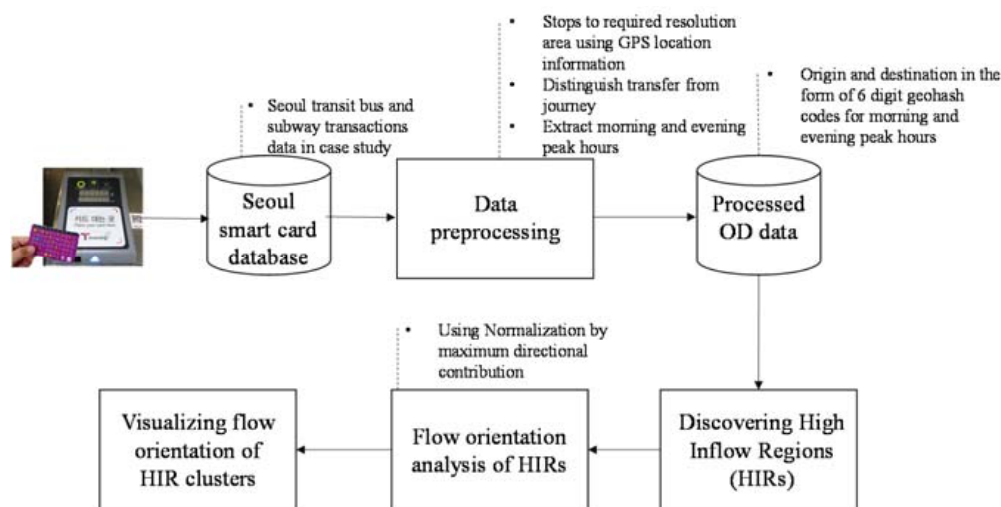


Figure 2. Flow orientation analysis of Seoul transportation data.

4.1. Data Pre-Processing

In this study, we employed the bus and subway transit data of Seoul, Korea. Weekday data for 2 February 2012, were used for the analysis. The attributes of the data used included passenger number, origin and destination station information, bus/rail type, and time stamp. These data were then transformed according to the proposed methodology so that the resulting origin and destination information was in the form of 6-digit Geohash codes. The size of a 6-digit Geohash region is approximately 1.2 km by 0.6 km in Korea, which depends upon the latitude of the region. Considering the distance between adjacent Geohash region, the area covered under 6-digit Geohash codes was found most suitable, because it was a walking distance and not large enough to use public transport for travelling within the Geohash region. In our OD data, 738 highlighted 6-digit Geohash regions existed. More specifically, they included a total of 736 destination regions and 737 origin regions as seen in the database.

In this study, we focused on the peak hours that could reflect major travel behavior on a day. A total of 4.5 million trips during the day were then divided into 24 hourly data based on the boarding time. We chose morning peak hours from 7:00 to 10:00 and evening peak hours from 16:00 to 20:00 to the extract HIRs of the morning and evening commutes. Morning HIRs had a high probability of being places where people travelled for daily activities, while evening HIRs had a high probability of being residential areas. Both the morning and evening peak commute data were analyzed separately after removing transfers to induce final destinations. Finally, the resulting OD dataset used in this experiment had 1.09 million trips for the morning and 1.30 million trips for the evening commute.

4.2. Discovery of HIRs

In this subsection, the morning and evening peak datasets were investigated separately to identify the HIRs where the majority of day and night activities were concentrated. The Pareto principle was also assumed that 80% of the output was produced by 20% of the input to deal with major flow orientation patterns. The inflow distribution of regional inflows in the morning and evening peak hours is shown in Figure 3. In the figure, HIRs were sorted in descending order of the number of inflows. It could be found that the top 23% of regions (175 of 730 regions) had charge of 80.7% of trips (0.88 million among 1.09 million trips) for morning commutes, and the top 30% of regions (219 of 731 regions) had charge of 80.0% of trips (1.04 million among 1.30 million trips) for evening commutes.

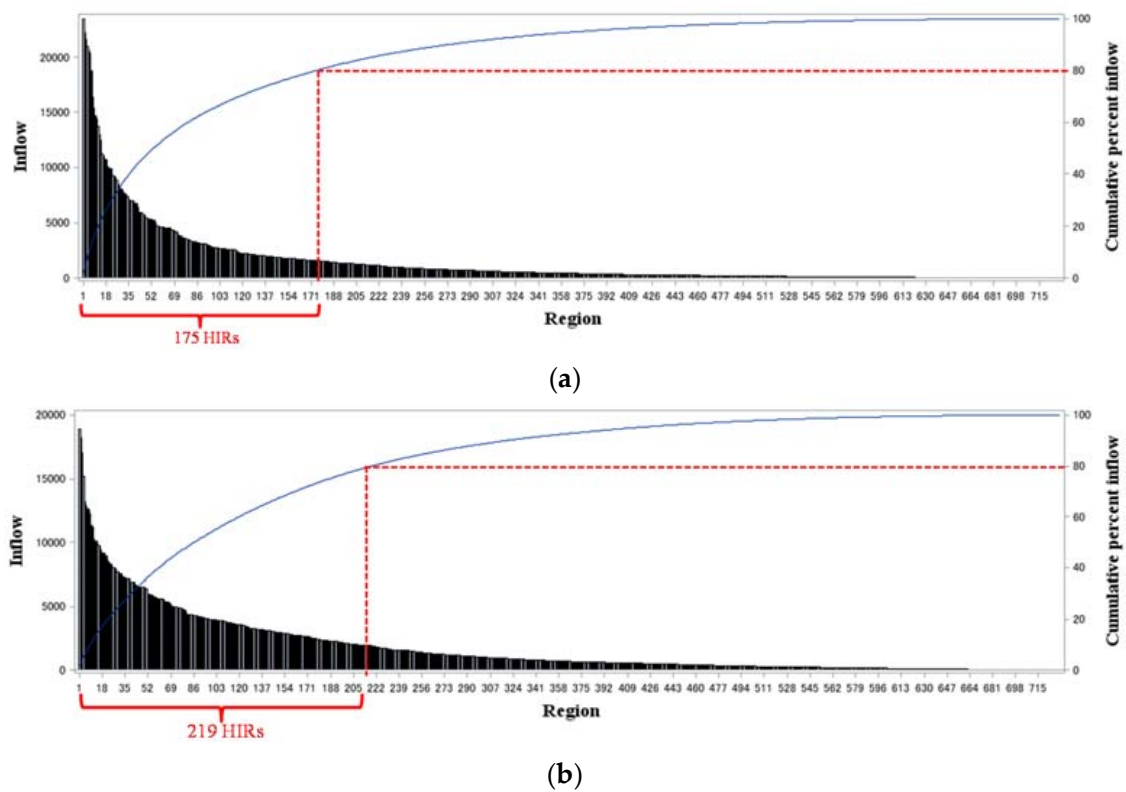


Figure 3. Distribution of inflows and selection of HIRs. (a) 175 HIRs selected from morning peak hour data; (b) 219 HIRs selected from evening peak commute data.

Figure 4 shows the locations of the 174 HIRs discovered from morning commute data and the 219 HIRs discovered from evening commute data on the city map. The identified HIRs for the morning commute, highlighted in blue, represent the day activity concentration areas, while the HIRs for the evening commute, highlighted in red, represent the night activity concentration areas, which are supposedly the residential areas. The HIRs in purple are the overlapped regions of the morning and evening HIRs.

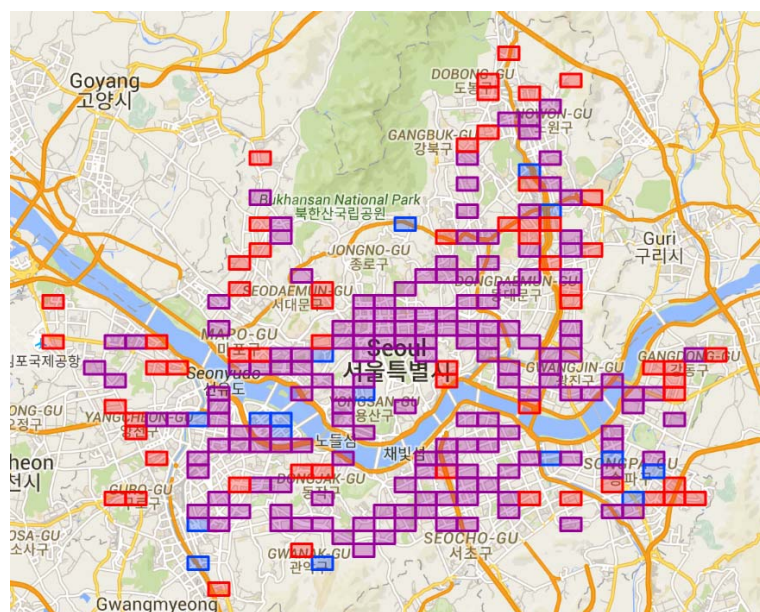


Figure 4. HIRs extracted in Seoul. 159 HIRs in purple both for morning and evening commutes, 16 HIRs in blue only for morning commute, and 59 HIRs in red only for evening commutes.

4.3. Flow Orientation Analysis

Once HIRs were selected, the analysis of flow orientation was performed for the HIRs. For each HIR, its origin regions were assigned to one of eight compass directions. Two directional inflow vectors F_i were then prepared for the morning and evening commutes by using Equation (3), and the dominance factor df_i was also calculated by using Equation (4). The results of the dominance factors and maximum directional inflows of the HIRs in Seoul are summarized with their Geohash codes in Table A1 of Appendix A.

The df values gave the impression at first glance that the particular HIR was symmetrically connected to all of the directions in terms of flow volume. For 175 morning commute HIRs in Seoul, the mean value of df was 0.542 and the standard deviation was 0.101. For 219 evening commute HIRs, the mean value of df was 0.614 and the standard deviation was 0.113. It could be said that residential and working places were concentrated in specific directions and, moreover, working places were more concentrated in certain directions than residences, since the evening commute HIRs had a higher df than the morning commute HIRs.

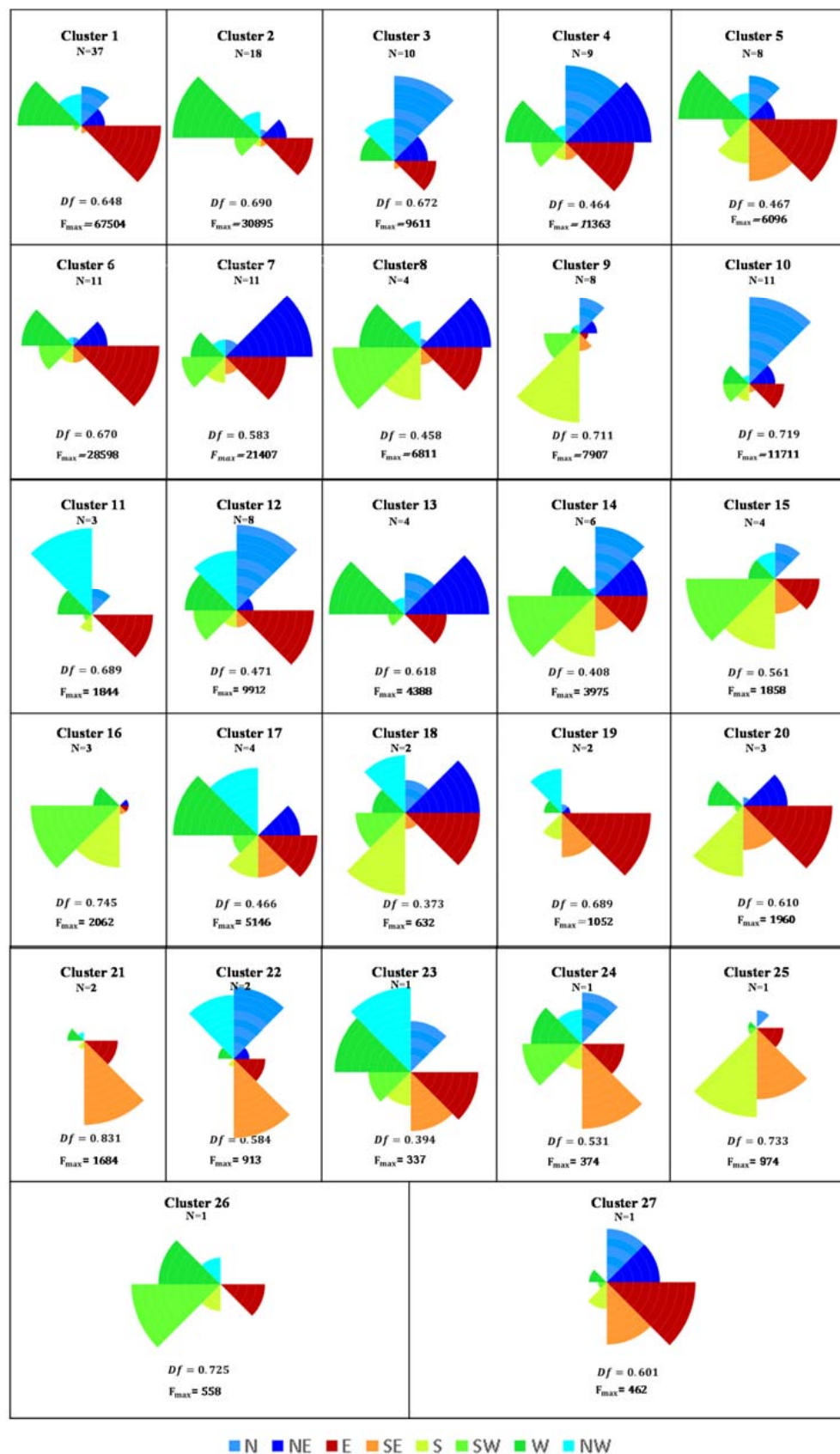
To examine flow orientation patterns, the clustering of HIRs was performed and the result was then visualized on the map. To achieve this, the AHC technique was applied in this study. AHC is an unsupervised bottom-up clustering method for making hierarchical groups of instances based on the similarity among instances [33]. AHC can generate a dendrogram as an output, which shows the progressive grouping of instances. The result gives the insight of dissimilarity in instances and sufficient options to select the suitable number of clusters. Hence, the AHC was adopted to understand the orientation flow patterns of HIRs in this research.

In this study, the similarity between the directional inflow vectors F_i of two HIRs was used to cluster the HIRs with AHC. Since F_i represents the normalized flow orientation from all eight directions for the HIR, clustering HIRs for F_i can find similar flow orientation patterns among HIRs. More specifically, average link clustering based on the Euclidean distance was opted. This clustering was justifiable in line with identifying HIRs having similar flow orientations. Figure A1 in Appendix B shows two dendrograms representing the progress of hierarchically clustering the HIRs for morning and evening commutes. Based on the evaluation measures of cubic clustering criteria, such as R-square, pseudo f, and pseudo T-square, 27 HIR clusters for the morning commute and 15 HIR clusters for the evening commute were identified. The results of HIR clustering for the morning and evening commutes were visualized with df and f^{max} as shown in Figure 5.

The morning HIRs shown in Figure 5a signify clustering of major business or study destinations and, conversely, the evening HIRs in Figure 5b signify major residential destinations. In Figure 5a,b, the HIRs with the same number or the same color belong to the same cluster, and their flow orientation patterns can be found in Figure 6a,b with the cluster number. For example, many HIRs in the morning commutes shown in Figure 5a were clustered in Cluster 1 in yellow, which directional orientation was mainly E and W directions as depicted in Figure 6a.

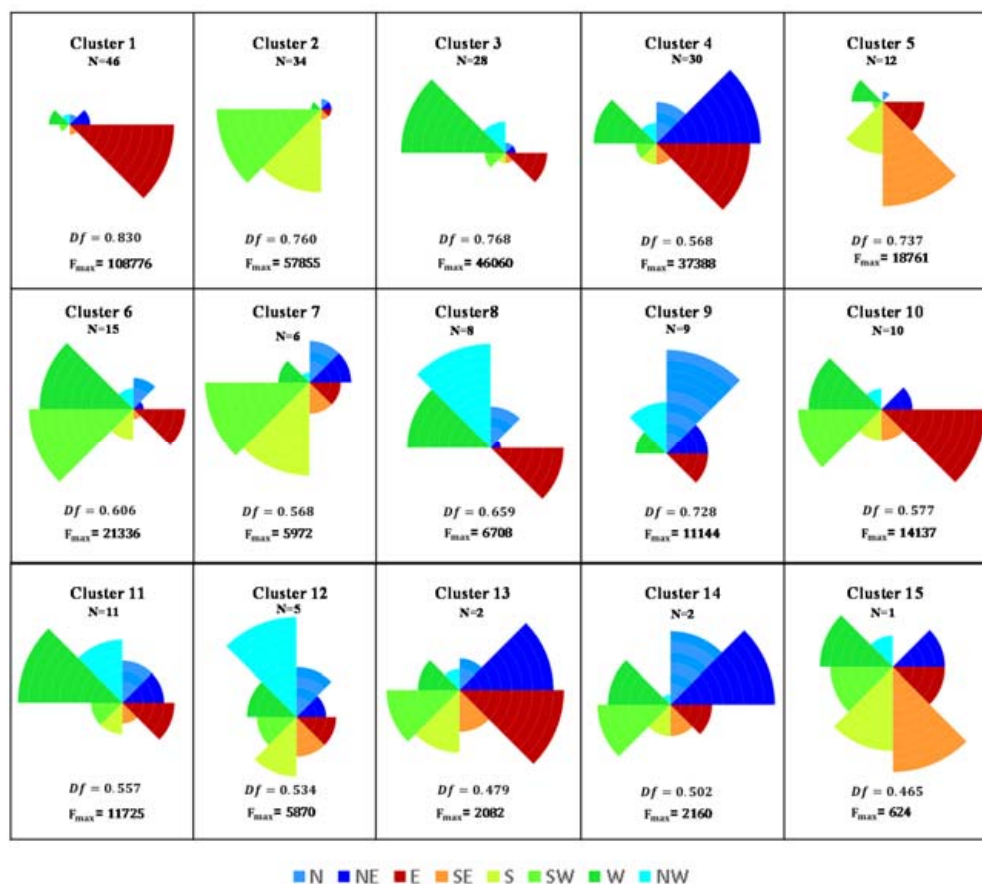
On the two maps with locations of HIR clusters, many of the near HIRs have similar flow orientation patterns for the morning and evening commutes. It is reasonable to believe that near regions have higher probabilities of similar structure of residential and working places. Nevertheless, it could be found that flow orientation similarity among near HIRs in the morning is less than the similarity in the evening by comparing Figure 5a,b. In other words, there were many similar flow orientation patterns among near locations in Figure 5b and also the orientation patterns were directed to the center of the city when inspecting the major directions shown in Figure 6b. However, morning orientation patterns shown in Figure 5a were more diverse than evening orientation patterns in Figure 5b.

Moreover, near regions are often segregated by transportation convenience, such as expressways and subway lines, and surrounding environments, such as rivers and parks. For example, most of the morning HIRs in the bottom of Figure 5a belong to Cluster 1 in yellow, which had mainly horizontal movement, such as the east and west directions (see the first fan diagram in Figure 6a), because the south area of Seoul is separated from the center and the north areas by the Han River. Likewise,



(a)

Figure 6. Cont.



(b)

Figure 6. Flow orientation patterns of HIR clusters. (a) Morning HIR clusters; (b) Evening HIR clusters.

In Figure 6a, we could find more highly clustered regions for the evening commute compared to those for the morning commute shown in Figure 5a. This result was interesting, since evening commute patterns were expected to be symmetrical to the morning commute patterns in that residential and working places were exchanged between the two commutes. It could be explained that many of the evening commutes still remained from 20:00 to 23:00 except for what we considered to be the evening commutes, 16:00–20:00. The evening commute HIRs were expected to contain many new trips in the evening. It is expected that a variety of information can be induced from the results of flow orientation analysis and that this information can be utilized for the purpose of citizen movement analysis and urban development.

5. Conclusions and Future Work

Using smartcard transit data, an analysis for discovering HIRs was used to understand their complexity by showing that daily activities were concentrated in limited areas. In this study, we identified morning commute HIRs where most of the day time activities were concentrated and evening commute HIRs that were supposed to be the residential regions. The geographical abstraction based on Geohash codes made it easy to represent and identify HIRs on a city map. Along with detecting HIRs, we also provided flow orientation analysis for each HIR in this study, which expanded the variation in the attracted orientation of HIR in terms of the eight compass directions. The dominance factor df presented the balance in flow orientation for any region at a glance, and flow orientation vectors were used for in-depth analysis of flow by applying classical data mining and statistical techniques. In this study, this vector was utilized to create HIR clusters that could visualize flow

orientation patterns in a city. The analysis illustrated the proposed methodology with the smart card transit data that had been collected from the subway and bus networks in Seoul. The results from the real data provided HIR clusters for flow orientation.

This methodology can also be applied to smart card data for other places to study existing scenarios and to point to exceptions in movement. Analyzing existing scenarios is a prerequisite for making operational changes or planning an upgradation. For understanding the cause of high or low inflows, analysis results should be referred to with land use information and network data. This process can help transport service providers when adopted before any operational changes or transit network improvements are made, and for urban planning. For other urban development projects, it is beneficial when combined with the socio-economic features of passengers travelling to destinations.

The approach to flow orientation analysis might be straightforward to analyze the movement patterns in a city. Therefore, the proposed flow orientation analysis can provide basic information on how people travel in public transportation networks according to time and visualize main movement in the city. It is helpful to understand major movements of people related to context such as daytime activities and residence. In practice, more detailed analysis on flow orientation and human movement could surely be conducted to induce the patterns according to various types of activities and different time periods such as weekday and weekend.

For future studies, this research could be extended to a longer period and a different scales of areas, although the analysis in this study was conducted using one-day data in a city. In future research, it would also be interesting to compare inflow and outflow patterns even though this study was only focused on incoming flow orientation patterns. In addition, the context information of a city, such as point-of-interest [20] and accessibility [24], could be integrated to provide more comprehensible results of flow orientation analysis results to civil planners.

Acknowledgments: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2013R1A2A2A03014718).

Author Contributions: P.S. conceived the experiments; J.-Y.J. designed the experiments; K.O. pre-processed the data; P.S. performed the experiments; K.O. visualized the results; P.S. and J.-Y.J. analyzed the data; P.S. and J.-Y.J. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A

Table A1. Description of HIRs in Seoul.

(a) Morning commute HIRs.

HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}
HIR1	wydm6d	0.63	8005	HIR51	wydm6e	0.69	2187	HIR101	wydmfx	0.63	897	HIR151	wydxg	0.71	777
HIR2	wydm9w	0.49	5460	HIR52	wydm6l	0.69	2128	HIR102	wydxr7	0.53	717	HIR152	wydxs8	0.38	367
HIR3	wydm9m	0.53	5782	HIR53	wydm1w	0.71	2244	HIR103	wydm2j	0.73	1205	HIR153	wydx5b	0.52	471
HIR4	wydm75	0.56	5983	HIR54	wydm3p	0.52	1370	HIR104	wydxqz	0.51	683	HIR154	wydm93	0.33	339
HIR5	wydm9x	0.50	5136	HIR55	wydx9	0.46	1149	HIR105	wydx59	0.66	964	HIR155	wydmgu	0.59	545
HIR6	wydm6f	0.55	5667	HIR56	wydmfb	0.32	862	HIR106	wydm69	0.68	1014	HIR156	wydmdb	0.34	337
HIR7	wydm7k	0.58	5546	HIR57	wydxrt	0.44	1049	HIR107	wydmfv	0.45	598	HIR157	wydxw6	0.59	532
HIR8	wydm9k	0.39	3310	HIR58	wydmee	0.62	1545	HIR108	wydxry	0.65	934	HIR158	wydmes	0.76	911
HIR9	wydm2n	0.56	4344	HIR59	wydmr	0.47	1102	HIR109	wydm7u	0.51	655	HIR159	wydxw	0.53	462
HIR10	wydm9y	0.66	5326	HIR60	wydxrx	0.36	903	HIR110	wydm1n	0.46	779	HIR160	wydmgr	0.49	417
HIR11	wydm8s	0.55	3928	HIR61	wydmf8	0.54	1246	HIR111	wydm9l	0.51	657	HIR161	wydxwt	0.63	583
HIR12	wydm9r	0.59	4202	HIR62	wydmmed	0.49	1109	HIR112	wydm1m	0.48	821	HIR162	wydmmp	0.47	515
HIR13	wydm9z	0.52	3350	HIR63	wydxq4e	0.61	1445	HIR113	wydm24	0.63	837	HIR163	wydm22	0.61	522
HIR14	wydm9v	0.46	2875	HIR64	wydmrsc	0.43	958	HIR114	wydmr	0.45	576	HIR164	wydmr	0.61	531
HIR15	wydm6x	0.42	2415	HIR65	wydxrv	0.58	1340	HIR115	wydx5m	0.72	1107	HIR165	wydx5x	0.68	632
HIR16	wydxpx	0.67	4167	HIR66	wydxrp	0.62	1464	HIR116	wydmr	0.34	566	HIR166	wydmr	0.46	374
HIR17	wydm60	0.62	3550	HIR67	wydm70	0.58	1318	HIR117	wydm6h	0.43	502	HIR167	wydm8u	0.60	516
HIR18	wydm6t	0.49	2630	HIR68	wydmr	0.50	1082	HIR118	wydx5n	0.68	874	HIR168	wydx56	0.53	438
HIR19	wydm8f	0.55	2794	HIR69	wydmr	0.58	1254	HIR119	wydmr	0.59	693	HIR169	wydmr	0.51	415
HIR20	wydm4x	0.29	2722	HIR70	wydm63	0.56	1191	HIR120	wydx5e	0.62	742	HIR170	wydmr	0.66	597
HIR21	wydm9t	0.56	2777	HIR71	wydmr	0.50	1035	HIR121	wydmr	0.37	537	HIR171	wydmr	0.58	484
HIR22	wydmr	0.36	1919	HIR72	wydmf7	0.43	850	HIR122	wydmr	0.59	635	HIR172	wydmr	0.38	558
HIR23	wydm7n	0.64	3223	HIR73	wydxr9	0.64	1300	HIR123	wydm9b	0.35	428	HIR173	wydxr8	0.61	520
HIR24	wydmr	0.38	1842	HIR74	wydm1z	0.54	1408	HIR124	wydmr	0.49	537	HIR174	wydmr	0.44	353
HIR25	wydmr	0.49	2189	HIR75	wydm89	0.64	1252	HIR125	wydxwm	0.53	568	HIR175	wydxrr	0.66	579
HIR26	wydmr	0.56	2529	HIR76	wydm6z	0.47	846	HIR126	wydmr	0.49	513				
HIR27	wydm6m	0.49	2045	HIR77	wydmr	0.53	929	HIR127	wydm87	0.53	580				
HIR28	wydm4r	0.40	2170	HIR78	wydm71	0.58	1039	HIR128	wydmr	0.36	414				
HIR29	wydmr	0.61	2625	HIR79	wydm0r	0.63	1162	HIR129	wydxrb	0.62	694				
HIR30	wydm6v	0.49	1965	HIR80	wydm2m	0.61	1093	HIR130	wydm0x	0.57	595				
HIR31	wydm65	0.47	1848	HIR81	wydxpr	0.40	886	HIR131	wydm3q	0.33	384				
HIR32	wydmr	0.43	1679	HIR82	wydmr	0.49	814	HIR132	wydmr	0.63	679				

Table A1. Cont.

(a) Morning commute HIRs.

HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}
HIR33	wydm9q	0.47	1767	HIR83	wydzj8	0.58	975	HIR133	wydm3h	0.57	592				
HIR34	wydm9n	0.56	2095	HIR84	wydm8h	0.46	754	HIR134	wydmk2	0.60	634				
HIR35	wydm2p	0.53	1885	HIR85	wydm1r	0.63	1077	HIR135	wydms6	0.62	636				
HIR36	wydm96	0.39	1442	HIR86	wydm3e	0.60	1005	HIR136	wydmkk	0.60	634				
HIR37	wydm2t	0.58	2050	HIR87	wydm3b	0.64	1100	HIR137	wydmk8	0.34	607				
HIR38	wydmdu	0.48	1680	HIR88	wydm8e	0.60	972	HIR138	wydm7s	0.71	857				
HIR39	wydmkm	0.63	2290	HIR89	wydm2r	0.67	1203	HIR139	wydwj5	0.65	1101				
HIR40	wydm8k	0.59	2068	HIR90	wydm8t	0.63	980	HIR140	wydmkg	0.63	655				
HIR41	wydm6k	0.60	2084	HIR91	wydmf4	0.65	1084	HIR141	wydm5z	0.46	575				
HIR42	wydm8n	0.66	2333	HIR92	wydm8g	0.50	763	HIR142	wydwjk	0.50	650				
HIR43	wydm8q	0.38	1192	HIR93	wydwq4	0.52	785	HIR143	wydm26	0.61	606				
HIR44	wydm8d	0.33	1101	HIR94	wydm3f	0.58	893	HIR144	wydmk9	0.70	777				
HIR45	wydm0z	0.57	1703	HIR95	wydm2c	0.61	947	HIR145	wydm8f	0.54	502				
HIR46	wydwq5	0.62	1849	HIR96	wydwq0	0.52	974	HIR146	wydm2q	0.58	556				
HIR47	wydwjr	0.43	1239	HIR97	wydwj2	0.69	1117	HIR147	wydmf6	0.66	681				
HIR48	wydm90	0.70	2275	HIR98	wydm8g	0.56	780	HIR148	wydmk7	0.58	549				
HIR49	wydm9j	0.58	1625	HIR99	wydm8f	0.38	551	HIR149	wydmcy	0.45	421				
HIR50	wydm85	0.53	1410	HIR100	wydm38	0.56	774	HIR150	wydm0u	0.49	588				

(b) Evening commute HIRs.

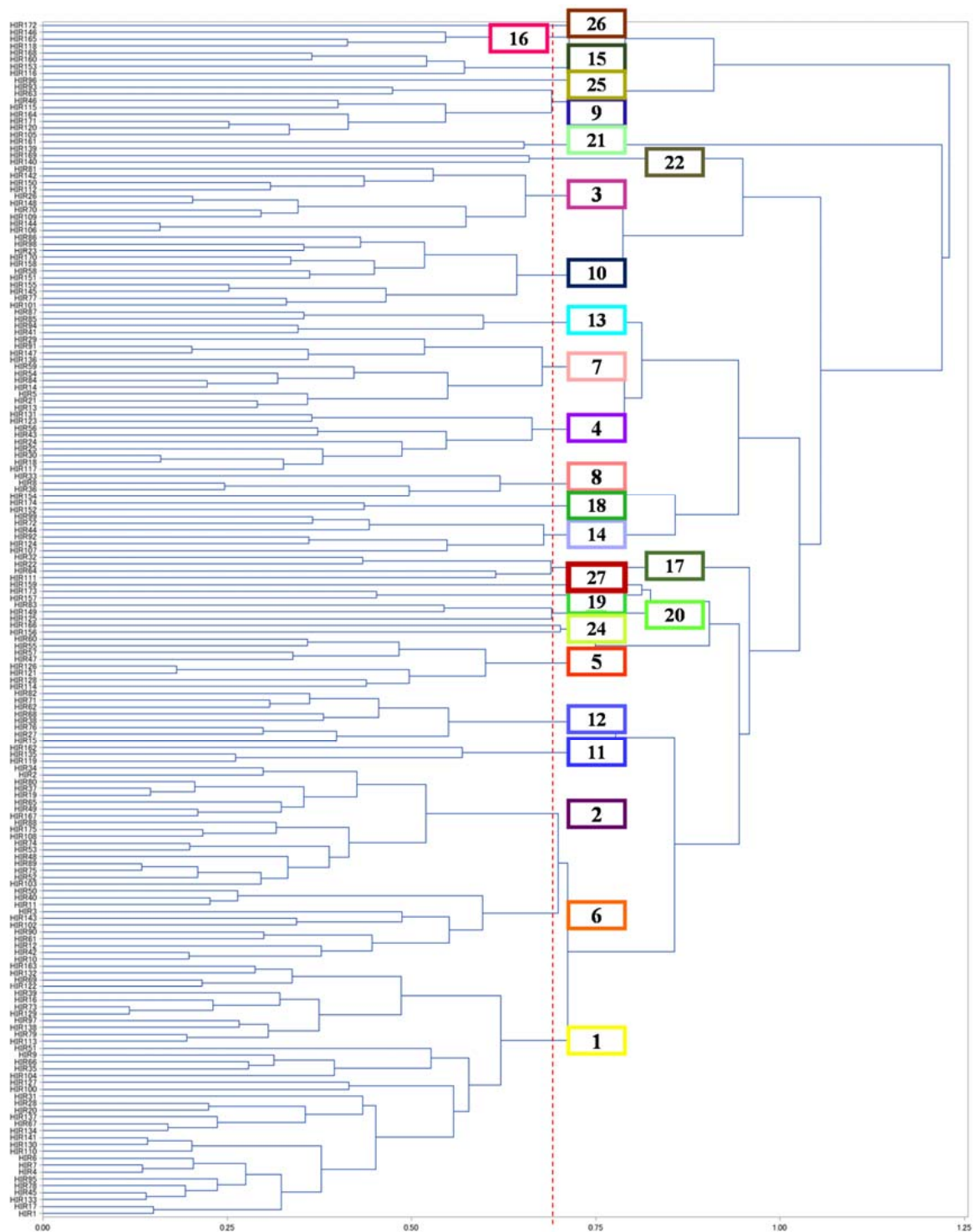
HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}
HIR1	wydm6d	0.62	6254	HIR51	wydm85	0.74	2992	HIR101	wydm8v	0.62	1266	HIR151	wydms8	0.52	756	HIR201	wydm8h	0.65	726
HIR2	wydm8s	0.70	7309	HIR52	wydwjr	0.71	2582	HIR102	wydm6h	0.40	813	HIR152	wydm3q	0.56	809	HIR202	wydm3f	0.64	705
HIR3	wydm0r	0.73	7758	HIR53	wydm8c	0.56	1673	HIR103	wydwjc	0.54	1066	HIR153	wydm8d	0.50	718	HIR203	wydm8u	0.51	516
HIR4	wydm9x	0.48	3529	HIR54	wydm8b	0.53	1546	HIR104	wydmmp	0.45	1150	HIR154	wydwq6	0.67	1710	HIR204	wydm8b	0.66	741
HIR5	wydm65	0.53	3425	HIR55	wydwq4	0.70	2447	HIR105	wydm1m	0.61	1862	HIR155	wydwjx	0.70	1194	HIR205	wydwj8	0.46	866
HIR6	wydm8d	0.57	3725	HIR56	wydm9r	0.56	1622	HIR106	wydmfv	0.70	1600	HIR156	wydm8m	0.75	1397	HIR206	wydm8c	0.61	650
HIR7	wydm8k	0.73	5886	HIR57	wydm96	0.40	1177	HIR107	wydm6k	0.44	868	HIR157	wydm3p	0.55	768	HIR207	wydwjx	0.74	953
HIR8	wydwq5	0.69	5031	HIR58	wydwq4	0.68	2185	HIR108	wydwj7	0.55	1067	HIR158	wydmf8	0.65	969	HIR208	wydm8d	0.40	411
HIR9	wydm0z	0.71	5183	HIR59	wydwjq	0.79	3416	HIR109	wydm8e	0.43	825	HIR159	wydm6t	0.33	515	HIR209	wydm63	0.63	666
HIR10	wydwq4	0.73	5196	HIR60	wydwq5	0.72	2496	HIR110	wydm2n	0.58	1111	HIR160	wydwjm	0.70	1123	HIR210	wydm87	0.66	725
HIR11	wydwjr	0.66	4078	HIR61	wydwq5	0.79	3236	HIR111	wydm7n	0.45	841	HIR161	wydwjg	0.74	1326	HIR211	wydwq4	0.75	947
HIR12	wydm9z	0.41	2052	HIR62	wydm8j	0.48	1333	HIR112	wydm2m	0.57	1075	HIR162	wydm7c	0.73	1256	HIR212	wydm8e	0.64	662
HIR13	wydm9k	0.55	2744	HIR63	wydm4x	0.29	1481	HIR113	wydwj5	0.78	2105	HIR163	wydm70	0.64	939	HIR213	wydm8e	0.66	717

Table A1. Cont.

(b) Evening commute HIRs.

HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}	HIR	Geohash	df	f^{max}
HIR14	wydmkm	0.70	4175	HIR64	wydm9m	0.59	1654	HIR114	wydm6v	0.52	949	HIR164	wydmfg	0.70	1129	HIR214	wydqh0	0.76	988
HIR15	wydq00	0.60	4414	HIR65	wydmg6	0.58	1596	HIR115	wydm89	0.66	1349	HIR165	wydm9j	0.68	1034	HIR215	wydm67	0.59	580
HIR16	wydmfx	0.63	3269	HIR66	wydmdb	0.62	1734	HIR116	wydmnb	0.63	1230	HIR166	wydm3h	0.58	788	HIR216	wydmk2	0.70	793
HIR17	wydm75	0.56	2595	HIR67	wydmdd	0.63	1756	HIR117	wydm9y	0.58	1040	HIR167	wydrjf	0.74	1274	HIR217	wydrjv	0.57	779
HIR18	wydmsc	0.63	3064	HIR68	wydm4r	0.53	1915	HIR118	wydm9v	0.74	1703	HIR168	wydm5b	0.76	1365	HIR218	wydmk6	0.60	590
HIR19	wydrjr	0.58	2682	HIR69	wydm9s	0.62	1676	HIR119	wydrjt	0.63	1197	HIR169	wydm91	0.48	624	HIR219	wydm69	0.61	574
HIR20	wydm7u	0.68	3561	HIR70	wydmddh	0.35	951	HIR120	wydm9m	0.65	1256	HIR170	wydm91	0.41	706				
HIR21	wydm1w	0.69	3595	HIR71	wydm9m	0.74	2371	HIR121	wydrxu	0.77	1921	HIR171	wydrz8	0.63	873				
HIR22	wydm6x	0.56	2355	HIR72	wydrx9	0.74	2346	HIR122	wydm9n	0.68	1374	HIR172	wydmf6	0.52	664				
HIR23	wydm7k	0.53	2226	HIR73	wydm9g	0.65	1778	HIR123	wydm2e	0.67	1283	HIR173	wydm9n	0.65	915				
HIR24	wydrpx	0.70	3378	HIR74	wydmkq	0.72	2142	HIR124	wydm9t	0.55	974	HIR174	wydr7c	0.44	1701				
HIR25	wydrjr	0.70	3342	HIR75	wydrwt	0.74	2353	HIR125	wydm9b	0.42	744	HIR175	wydrx8	0.75	1240				
HIR26	wydm9w	0.44	1766	HIR76	wydm26	0.70	2028	HIR126	wydm2c	0.66	1256	HIR176	wydm93	0.22	754				
HIR27	wydm9s	0.62	2583	HIR77	wydrq9	0.68	1882	HIR127	wydm9y	0.64	1143	HIR177	wydrq5	0.62	1186				
HIR28	wydm2t	0.57	2278	HIR78	wydm9f	0.51	1220	HIR128	wydrx5	0.75	1657	HIR178	wydm0t	0.61	758				
HIR29	wydrw2	0.81	5148	HIR79	wydm9e	0.74	2314	HIR129	wydm9s	0.72	1431	HIR179	wydm6z	0.40	469				
HIR30	wydm9t	0.43	1661	HIR80	wydm9u	0.53	1678	HIR130	wydm9s	0.61	1027	HIR180	wydm9b	0.60	743				
HIR31	wydmf7	0.68	2941	HIR81	wydm9l	0.74	2110	HIR131	wydm9w	0.66	1200	HIR181	wydm9e	0.49	570				
HIR32	wydm1n	0.72	3282	HIR82	wydm73	0.54	1177	HIR132	wydrz5	0.34	738	HIR182	wydm61	0.67	889				
HIR33	wydm9t	0.60	2228	HIR83	wydm9v	0.66	1590	HIR133	wydrq7	0.71	2373	HIR183	wydm9k	0.49	570				
HIR34	wydmcc	0.50	1795	HIR84	wydm1z	0.44	1243	HIR134	wydm9d	0.46	734	HIR184	wydm9c	0.71	1000				
HIR35	wydrq5	0.76	3768	HIR85	wydrq5	0.77	2306	HIR135	wydm9k	0.54	1170	HIR185	wydrjt	0.64	1220				
HIR36	wydm60	0.52	1846	HIR86	wydm93	0.69	1733	HIR136	wydrqx	0.82	2223	HIR186	wydm9k	0.67	871				
HIR37	wydm38	0.63	2385	HIR87	wydrq5	0.70	1787	HIR137	wydm9g	0.70	1276	HIR187	wydm3b	0.64	788				
HIR38	wydmft	0.69	2808	HIR88	wydm9f	0.64	1471	HIR138	wydm9h	0.62	1028	HIR188	wydrjr	0.56	631				
HIR39	wydmec	0.58	2015	HIR89	wydrjr	0.66	1535	HIR139	wydm9w	0.70	1297	HIR189	wydm9y	0.58	672				
HIR40	wydm6m	0.47	1611	HIR90	wydm9v	0.37	811	HIR140	wydm9t	0.65	1112	HIR190	wydm96	0.53	802				
HIR41	wydm9q	0.71	2937	HIR91	wydm9n	0.62	1364	HIR141	wydm92	0.58	893	HIR191	wydm9z	0.41	802				
HIR42	wydm8f	0.49	1664	HIR92	wydm9g	0.53	1105	HIR142	wydm94	0.66	1117	HIR192	wydrq5	0.75	1095				
HIR43	wydrw6	0.79	3989	HIR93	wydm9d	0.67	1512	HIR143	wydrq3	0.76	1624	HIR193	wydm94	0.58	658				
HIR44	wydmfb	0.69	2624	HIR94	wydm90	0.49	983	HIR144	wydm9t	0.52	1076	HIR194	wydm9m	0.50	537				
HIR45	wydm9p	0.53	1695	HIR95	wydrq7	0.81	2711	HIR145	wydm9k	0.60	943	HIR195	wydrjp	0.58	892				
HIR46	wydm9r	0.72	2907	HIR96	wydm95	0.54	1102	HIR146	wydm9c	0.41	624	HIR196	wydm9v	0.71	903				
HIR47	wydm0x	0.73	3046	HIR97	wydrjr	0.67	1482	HIR147	wydm9c	0.61	939	HIR197	wydm9q	0.41	435				
HIR48	wydm6f	0.50	1591	HIR98	wydm9k	0.72	1750	HIR148	wydm9e	0.55	822	HIR198	wydm9g	0.73	989				
HIR49	wydm1r	0.63	2133	HIR99	wydrqw	0.79	2354	HIR149	wydrqu	0.73	1358	HIR199	wydrq6	0.77	1125				
HIR50	wydrjp	0.75	3207	HIR100	wydm91	.56	1116	HIR150	wydm9j	0.58	870	HIR200	wydm9n	0.75	1016				

Appendix B



(a)

Figure A1. Cont.

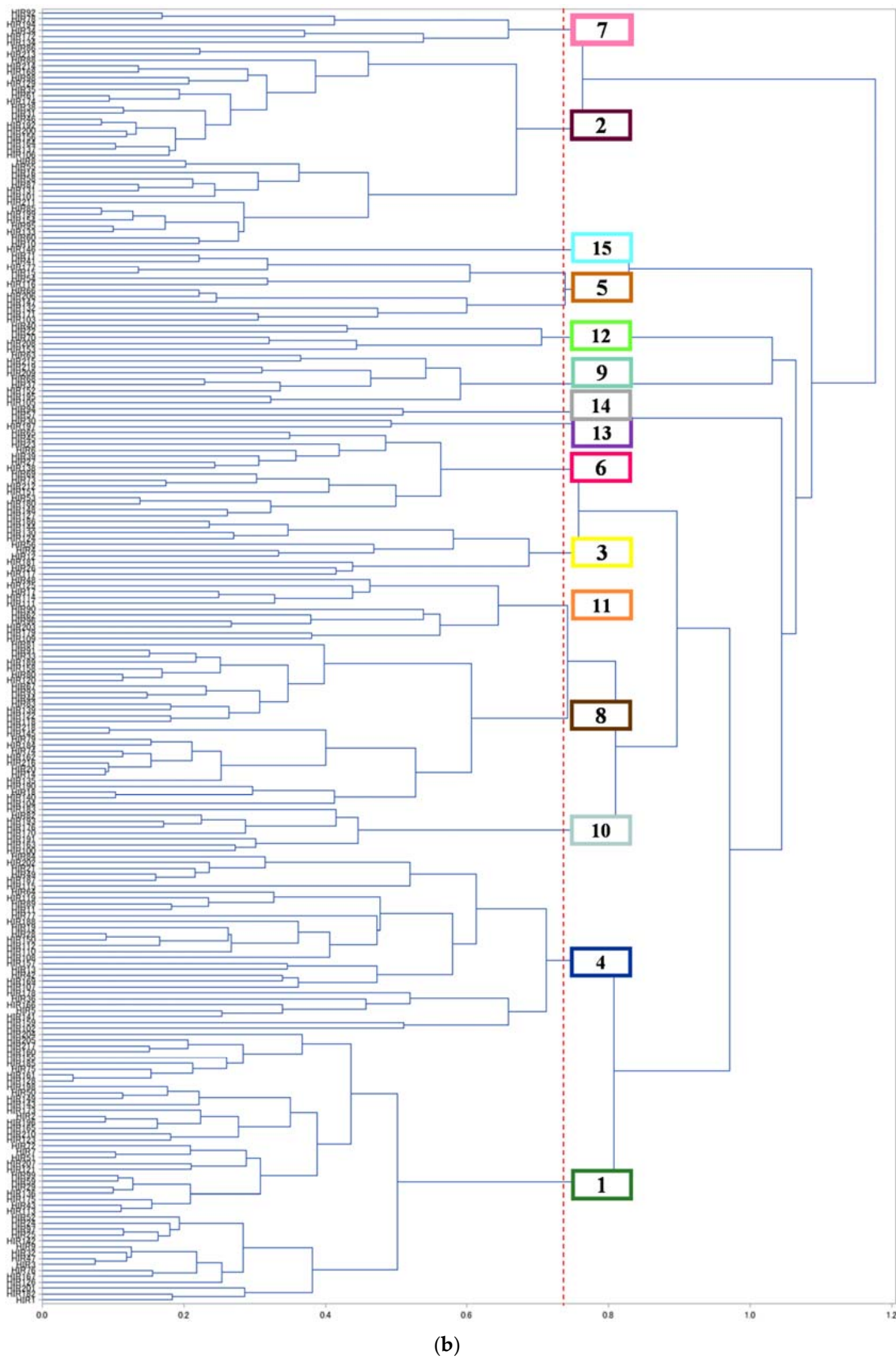


Figure A1. Dendograms of HIR clustering. (a) Morning commute HIRs; (b) Evening commute HIRs.

References

1. McMillen, D.P.; McDonald, J.F. A nonparametric analysis of employment density in a polycentric city. *J. Reg. Sci.* **1997**, *37*, 591–612. [CrossRef]
2. Jun, M.J.; Ha, S.K. Evolution of employment centers in Seoul. *Rev. Urban Reg. Dev. Stud.* **2002**, *14*, 117–132. [CrossRef]
3. Baumont, C.; Ertur, C.; Gallo, J. Spatial analysis of employment and population density: the case of the agglomeration of Dijon 1999. *Geogr. Anal.* **2004**, *36*, 146–176. [CrossRef]
4. Roth, C.; Kang, S.M.; Batty, M.; Barthélemy, M. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS ONE* **2011**, *6*, e15923. [CrossRef] [PubMed]
5. Zhong, C.; Arisona, S.M.; Huang, X.; Batty, M.; Schmitt, G. Detecting the dynamics of urban structure through spatial network analysis. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 2178–2199. [CrossRef]
6. Craig, S.G.; Kohlhasse, J.E.; Perdue, A.W. Empirical polycentricity: The complex relationship between employment centers. *J. Reg. Sci.* **2016**, *56*, 25–52. [CrossRef]
7. Yang, X.; Fang, Z.; Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Lin, Y. Understanding spatiotemporal patterns of human convergence and divergence using mobile phone location data. *ISPRS Int. J. Geo-Inf.* **2016**. [CrossRef]
8. Helsley, R.W.; Sullivan, A.M. Urban subcenter formation. *Reg. Sci. Urban Econ.* **1991**, *21*, 255–275. [CrossRef]
9. Geohash. *Geohash Tips & Tricks*; 2017. Available online: <http://Geohash.org/site/tips.html> (accessed on 1 May 2017).
10. Pelletier, M.P.; Trépanier, M.; Morency, C. Smart card data use in public transit: A literature review. *Transp. Res. C Emer. Technol.* **2011**, *19*, 557–568. [CrossRef]
11. Morency, C.; Trépanier, M.; Agard, B. Analysing the Variability of Transit Users' Behaviour with Smart Card Data. In Proceedings of the 19th International IEEE Intelligent Transportation Systems Conference (ITSC), Toronto, ON, Canada, 17–20 September 2006; pp. 44–49.
12. Morency, C.; Trépanier, M.; Agard, B. Measuring transit use variability with smart-card data. *Transp. Policy* **2007**, *14*, 193–203. [CrossRef]
13. Ma, X.; Wu, Y.J.; Wang, Y.; Chen, F.; Liu, J. Mining smart card data for transit riders' travel patterns. *Transp. Res. C Emer. Technol.* **2013**, *36*, 1–12. [CrossRef]
14. Kieu, L.M.; Bhaskar, A.; Chung, E. Passenger segmentation using smart card data. *IEEE Trans. Intell. Transp.* **2015**, *16*, 1537–1548. [CrossRef]
15. Kim, K.; Oh, K.; Lee, Y.K.; Kim, S.; Jung, J.-Y. An analysis on movement patterns between zones using smart card data in subway networks. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1781–1801. [CrossRef]
16. Du, B.; Yang, Y.; Lv, W. Understand Group Travel Behaviors in an Urban Area Using Mobility Pattern Mining. In Proceedings of the 10th IEEE International Conference on Ubiquitous Intelligence and Computing and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), Washington, DC, USA, 3–6 December 2013; pp. 127–133.
17. Tao, S.; Corcoran, J.; Mateo-Babiano, I.; Rohde, D. Exploring Bus Rapid Transit passenger travel behaviour using big data. *App. Geogr.* **2014**, *53*, 90–104. [CrossRef]
18. Tao, S.; Rohde, D.; Corcoran, J. Examining the spatial-temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *J. Transp. Geogr.* **2014**, *41*, 21–36. [CrossRef]
19. Long, Y.; Thill, J.C. Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. *Comp. Environ. Urban Syst.* **2015**, *53*, 19–35. [CrossRef]
20. Zeng, W.; Fu, C.W.; Arisona, S.M.; Schubiger, S.; Burkhard, R.; Ma, K.L. Visualizing the Relationship Between Human Mobility and Points of Interest. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2271–2284. [CrossRef]
21. Zhong, C.; Huang, X.; Arisona, S.M.; Schmitt, G.; Batty, M. Inferring building functions from a probabilistic model using public transportation data. *Comput. Environ. Urban Syst.* **2014**, *48*, 124–137. [CrossRef]
22. Bagchi, M.; White, P.R. The potential of public transport smart card data. *Transp. Policy* **2005**, *12*, 464–474. [CrossRef]
23. Kusakabe, T.; Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transp. Res. C Emer Technol.* **2014**, *46*, 179–191. [CrossRef]
24. Cats, O.; Wang, Q.; Zhao, Y. Identification and classification of public transport activity centres in Stockholm using passenger flows data. *J. Transp. Geogr.* **2015**, *48*, 10–22. [CrossRef]

25. Zhu, X.; Guo, D. Mapping large spatial flow data with hierarchical clustering. *Trans. GIS* **2014**, *18*, 421–435. [[CrossRef](#)]
26. Song, Y.; Lee, K.; Anderson, W.P.; Lakshmanan, T.R. Industrial agglomeration and transport accessibility in metropolitan Seoul. *J. Geogr. Syst.* **2012**, *14*, 299–318. [[CrossRef](#)]
27. Wu, W.; Xu, J.; Zeng, H.; Zheng, Y.; Qu, H.; Ni, B.; Ni, L.M. Telcovis: Visual exploration of co-occurrence in urban human mobility based on telco data. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 935–944. [[CrossRef](#)] [[PubMed](#)]
28. Andrienko, G.; Andrienko, N.; Hurter, C.; Rinzivillo, S.; Wrobel, S. Scalable analysis of movement data for extracting and exploring significant places. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 1078–1094. [[CrossRef](#)] [[PubMed](#)]
29. Bahamonde, J.; Hevia, A.; Font, G.; Bustos-Jimenez, J.; Montero, C. Mining private information from public data: The Transantiago Case. *IEEE Pervas. Comp.* **2014**, *13*, 37–43. [[CrossRef](#)]
30. Ma, Y.; Xu, W.; Zhao, X.; Li, Y. Modeling the hourly distribution of population at a high spatiotemporal resolution using subway smart card data: A case study in the central area of Beijing. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 128. [[CrossRef](#)]
31. Sanders, R. The Pareto principle: Its use and abuse. *J. Serv. Mark.* **1987**, *1*, 37–40. [[CrossRef](#)]
32. Juran, J.M.; Gryna, F.M. *Juran's Quality Control Handbook*, 5th ed.; McGraw-Hill: New York, NY, USA, 1998; ISBN 0-07-034003-X.
33. Olson, C.F. Parallel algorithms for hierarchical clustering. *Parallel Comput.* **1995**, *21*, 1313–1325. [[CrossRef](#)]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).