International Journal of
Geo-Information

MDPI

Article

Inferring Social Functions Available in the Metro Station Area from Passengers' Staying Activities in Smart Card Data

Yang Zhou [1,2,*], Zhixiang Fang [3], Qingming Zhan [4], Yaping Huang [2] and Xiongwu Fu [1]

[1] Wuhan Land Use and Urban Spatial Planning Research Center, 55 Sanyang Road, Wuhan 430014, China; fuxiongwu@wlsp.org.cn

[2] School of Architecture and Urban Planning, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan 430074, China; huangyaping@hust.edu.cn

[3] State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; zxfang@whu.edu.cn

[4] School of Urban Design, Wuhan University, 8 Donghu Nan Road, Wuhan 430072, China; qmzhan@whu.edu.cn

* Correspondence: cleverzhouyang@whu.edu.cn; Tel.: +86-27-6877-5699

Received: 13 October 2017; Accepted: 26 November 2017; Published: 1 December 2017

Abstract: The function of a metro station area is vital for city planners to consider when establishing a context-aware Transit-Oriented Development policy around the station area. However, the functions of metro station areas are hard to infer using the static land use distribution and other traditional survey datasets. In this paper, we propose a method to infer the functions occurring around the metro station catchment areas according to the patterns of staying activities derived from smart card data. We first define the staying activities by the spatial and temporal constraints of the two consecutive alighting and boarding records from the individual travel profile. Then we cluster and label the whole staying activities by considering the features of duration, frequency, and start time. By analyzing the percentage of different types of aggregated activities happening around each metro station, we cluster and explore the functions of the metro station area. Taking Wuhan as a case study, we analyze the results of Wuhan metro systems and discuss the similarities and differences between the functions and the land use distribution around the station area. The results show that although there exist some agreements, there is also a gap between the human activities and the land uses around the station area. These findings could give us deeper insight into how people act around the stations by metro systems, which will ultimately benefit the urban planning and policy development.

Keywords: social function; metro station areas; smart card data; staying activities; Transit-Oriented Development

1. Introduction

The Transit-Oriented Development (TOD) principle has been widely accepted and implemented when designing the transit surrounding areas in urban cities ever since Calthorpe [1] codified the concept of TOD. TOD aims to create a high density, a high diversity, and a design-oriented neighborhood within walking distance of public transportation in order to reduce private car use. However, it is strongly suggested that the TOD strategies be varied according to the intrinsic features of stations, such as form, land use mix, location, accessibility, and so on [2,3]. Among the indicators of TOD typology, the functional mix of a station area is an important input measurement to depict the development characteristics of TOD [4]. In this context, the social functions of station areas need to be examined comprehensively in order to help city planners gain insight into the existing conditions of human movement so as to make targeted and sustainable design decisions when redeveloping metro catchment areas based on the TOD expectations.

The social function of a metro station area is its socioeconomic orientation within the urban space with regards to the urban structural hierarchy. Where people reside, work, entertain, study, and socialize are mutually determined by the land use distribution in the station catchment areas and the individual travel pattern. Inferring the social function of the catchment area of a metro station directly from its surrounding land use pattern is difficult because (1) the static station's catchment area is not easily determined and (2) single land use may display multiple functions [5]. To interpret how a metro station area's function beyond the spatial distributions of its physical environment is still challenging.

With the support of geospatial big data, the socioeconomic environment of a place can be characterized by spatiotemporal mobility patterns and human activity features at different scales. Liu et al. [6] used the term social sensing to show that big geospatial data provides an alternative approach to uncover land uses and explore how cities function at a fine spatial and temporal resolution. During the past decade, there has been a significant amount of research work focusing on this topic using different types of data, such as taxi GPS data [7,8], mobile phone data [9], social media check-in data [10], etc. These studies enhanced our understanding about how the urban space functions from the perspective of human mobility pattern.

As a benefit of Automated Fare Collection (AFC) systems adopted by public transit agencies in many cities, a whole set of travelers' detailed check in and check out behavior in the transit network are tracked and stored as Smart Card Data (SCD). Such a data source provides rich information about people's travel habits and behavioral activity patterns when using public transport on a daily basis. By reconstructing all the check-in and check-out records from the travel log chronologically, the moving activities that happened during transit as well as the staying activities that occurred around the transit stations could be recognized. The moving activity is the trip that a traveler takes between the check-in and check-out actions. The staying activity is an anchoring duration that starts when a traveler checks out of the station and lasts until he/she boards again to make another trip. Transportation engineers and data mining analysts draw on smart card data to investigate the traveling behavior, including the moving stability [11], abnormality [12], temporally or spatially regularity [13,14], variability in regularity [15], and so forth. Urban planners and researchers study the mobility patterns in the urban fabric to understand the urban structure [16] and commuting patterns [17]. Existing studies take full advantage of the temporal changes in the aggregation of the check-in and check-out frequency; nevertheless, the anchoring durations of the staying activities have been somewhat ignored.

This paper aims to infer the social functions–occurring in metro station areas, considering the patterns of passengers' staying activities derived from smart card data. By demarcating the staying activities from individual travel profiles, we present a method to cluster different types of staying activities according to their start time, duration, and frequency. Then we distinguish and study the functions in the metro station catchment areas with regards to the composition of the activities that people perform around the stations.

Our contribution is twofold. First, we propose an easily implemented method to partition the transit records according to the staying behavior, which shows superiority in depicting the aggregated activity characteristics happening around the metro stations. Second, the metro station area functions in the urban behavioral space using staying activity patterns are observed and analyzed. Compared to the static land use distribution, the social function of the station areas revealed in this study has a strong connection to human movement patterns. The results are valuable for analyzing the relationships between the function and the organization as well as distribution of land use in metro station surrounding areas, which will finally benefit urban planning and public transportation management.

The remainder of this paper is organized as follows. Section 2 reviews the related work in smart card data. Section 3 briefly describes the study area and smart card preparation. Section 4 presents the methodology of the identification and classification of staying activities that happened around each metro station. In Section 5 we present a case study in Wuhan. Section 6 draws our conclusions and discusses future work.

2. Related Work

Ever since Bagchi and White [18] discussed the potential role of smart card data as a new source of travel behavior collection, an increasing amount of work has been done to serve transportation management and planning in the past two decades. Nassir et al. and Alsger et al. [19,20] used SCD to infer origin-destination matrices, which is a costly job using traditional travel survey data. Kusakabe & Asakura [21] estimated the trip purpose via the Bayes probabilistic model combined with travel survey data. Ma et al., Mahrsi et al., and Kieu et al. [13,22,23] measured passengers' travel habits and regularity over long periods of time to cluster cardholders. Zhou et al. [24] monitored and quantified the elasticity of distance travelled as transit feasibility in transit-served areas. Ma et al., Long et al., Long & Thill, and Zhou et al. [25–27] studied the commuting patterns anchored in the public transportation fabric.

In recent years, urban researchers used the smart card data as a large sampling population at a finer temporal and spatial granularity, taking advantage of the individual's daily movements in transit network recorded in SCD. Liu et al. [28] studied the movement patterns at both an individual and a collective scale. Lathia et al. [29] tested the relationship between urban mobility revealed by smart card data and social deprivation. Tao et al. [30] extracted and visualized aggregated flow patterns of bus passengers. Zhong et al. [15] explored how the variations in regularity scale at different temporal resolutions across different cities. These works provide us with new tools to better understand the dynamics of the spatial temporal distribution of human movement.

The urban structure is closely related to the subway systems, which can be considered in terms of urban interaction. For example, Roth et al. [16] studied the urban hierarchical structure using metro SCD in London. Urban morphology such as borders, hubs, and centers are also studied in [31,32]. Cats et al. [33] identified and classified urban activity centers according to the spatial proximity and travel flows of public transportation. Zhong et al. [34] inferred building functions from public smart data combined with survey data. Kim et al. [35] identified the movement patterns between spatially adjacent zones to understand the relationships between city areas. This study uses human tracking data to demonstrate the relationships between human activities and the urban functional environment.

Of the existing studies, it is common for researchers to investigate the temporal rhythms among all metro stations by mining the variations of check-in and check-out frequency in different time slots. If we divide a day into 24 h, 48 variables are needed to depict the check-in and check-out frequencies of a station in a day. The high dimension of the matrix makes the behavioral patterns dubiously and difficult to explain. For instance, Gong et al. [36] employed Eigen decomposition method to reduce the dimension in order to capture the common patterns of the passengers' variation over time. Nevertheless, station area functions other than the dominant ones (work and residences) are hard to infer because the passenger activity characteristics are not distinguished through card swiping behavior. Passenger activity patterns are generally derived from superficial card swiping behavior, so how to infer all station area functions by these patterns needs further investigation.

Whereas swiping frequency changes are associated with travelers' activity intensity and activity type, it is possible to present the characteristics of the metro station through patterns of aggregated staying activities that occurred around metro stations. For example, if a person checks out of the metro station at 8 a.m. and checks in at the same station at 6 p.m., his staying activity around the station has a high possibility to be identified as work. Following this logic, we partition the transit records according to the staying behavior that happened around the station area and investigate the aggregated activity patterns. Instead of centering on the trip episodes as in the existing literature, we spotlight the staying duration episodes to unveil the activity patterns hidden behind the card swiping behavior in this paper.

The concept of staying activity is not new, although it is addressed by different names. Chakirov et al. [37] denoted the activities between the trip chain as the consistent PT activities and targeted detecting home and work activities; Bouman et al. [38] identified the typical durations of the activity intervals. However, both works focused more on the duration distribution of the activity chains of individual passengers. The geographical environment around different stations has not yet been linked to human activity patterns. Furthermore, whether there are spatial differences of staying activities in different metro stations and how to infer the functions of the catchment area of a metro station from activities recorded in SCD remains unknown. To address this issue, this paper studies how the grouped staying activities in different station areas derived from SCD contribute to characterizing the social functions of the stations.

3. Study Area and Data Preparation

With a population of 8 million, Wuhan is the largest city in central China. As shown in Figure 1, the area inside the 3rd express ring road is the core area in which most of the inner city movement happens. After building its first subway line (Line 1) in 2004, Wuhan experienced rapid development. By 2049, Wuhan is expected to have built a well-connected subway network covering 1000 km across the whole metropolitan development area according to its long-term strategy plan [39]. As of August 2016, as shown in Figure 1, there are four operational subway lines, most of which are located inside the 3rd express ring road and have a total distance of 126 km. There are 96 metro stations in total, including six transit stations. The transit ridership in the metro systems accounted for 27.2% of all public transit in 2016 [40].

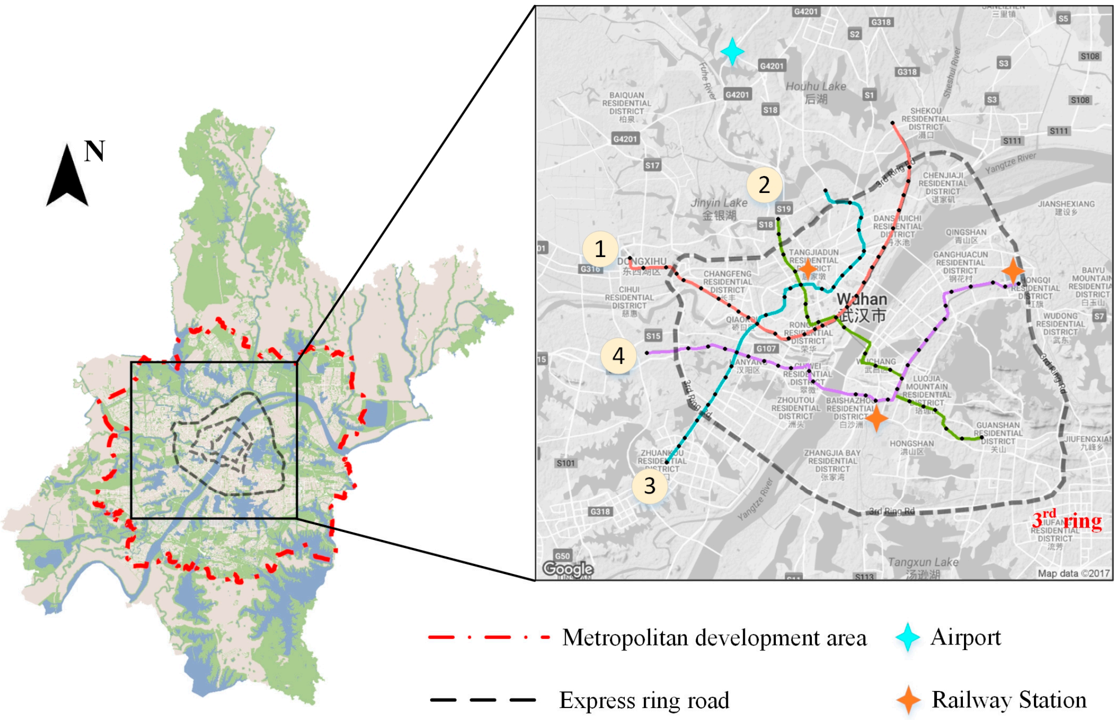


Figure 1. Metro lines in Wuhan. The area inside the 3rd express ring road is the core urban area of Wuhan.

In this study, we use the SCD dataset, which contains around 84.7 million transactions made by 5.6 million cardholders using the metro network from 1 to 31 August 2016. The original format included card ID, trade station ID, trade type (check-in or check-out), and trade time. We reconstruct the trip that each traveler took by connecting two successive check-in (boarding) and check-out (alighting) records for the same day, making sure that the travel time is shorter than 1.5 h, which is the maximum travel time allowed by Wuhan metro operators. Table 1 shows some examples of the reconstructed trip. The analysis in this paper focuses only on workdays (Monday to Friday).

Table 1. The format of reconstructed trip data.

Card ID	Boarding Station	Boarding Time	Alighting Station	Alighting Time
*****851	239	2016-08-02 14:15:36	448	2016-08-02 14:39:30
*****851	448	2016-08-02 21:28:10	239	2016-08-02 21:57:22
*****647	249	2016-08-30 08:37:47	247	2016-08-30 08:44:51
*****865	239	2016-08-24 10:09:50	246	2016-08-24 10:31:54

4. Methodology

Firstly, we bring forward the definition of staying activities by using both a temporal and a distance constraint. Then, by adopting a DBSCAN method, we aggregate and label the typical staying activities according to their start time, end time, and frequency. Finally, the station area functions are inferred from the proportion of different types of activities that happened around the station area using a *k*-means method.

4.1. Definition of Staying Activities

By querying all trip records that a passenger made and sorting by the boarding time through the dataset, we get every trip that he/she took chronologically. The trip that a passenger takes are denoted as trip_i:

$$\text{trip}_i = (s_{b,i}, s_{a,i}, t_{b,i}, t_{a,i}), \tag{1}$$

where $s_{b,i}$, $s_{a,i}$ are boarding and alighting stations; $t_{b,i}$, $t_{a,i}$ are the boarding and alighting time, respectively.

Generally speaking, a passenger takes the metro system to arrive at his or her destination to perform various activities, such as shopping, working, staying at home, etc. If he or she boards at the same station for another trip after finishing their activity, then a staying activity could be identified through the smart card data. When the passenger's destination is located in the middle of the two metro stations, he or she would probably choose to board at the nearest station. Hence, as illustrated in Figure 2, we are able to identify staying activity when the alighting station and the successive boarding station are located within a certain distance. Another constraint would be the staying duration. We need to filter those durations that last across days, because in this circumstance, there is a high possibility that the traveler will transfer to other transit modalities, like taxi, bus, or private car after performing the activity around the alighting station. The staying activity addressed here is similar to the consistent PT activity proposed in [37], except that only a distance constraint is adopted in consistent PT activity because the authors mainly aimed at the detection of home and work activities and the data used in [37] contained both bus and metro data, which had more consistency than the metro data in our study. Here we adopt the new terminology "staying activity" to highlight the staying geographical location of the activity.

According to the above discussion, considering two consecutive trips trip_i and trip_{i+1} made by a passenger, if $\text{dist}(s_{a,i}, s_{b,i+1}) \leqslant \delta_d$ and $t_{b,i+1} - t_{a,i} \leqslant \delta_t$, then a staying activity a_m that occurred around the station $s_{activity}$ is detected, see Equation (2):

$$\text{a}_m = (t_{start}, t_{end}, t_{duration}, s_{activity}) \tag{2}$$

where $s_{activity}$ denotes the metro station where the staying activity took place. We set $s_{activity} = s_{a,i}$, so the activity is always assigned to the alighting station when the alighting and boarding stations are different but within the distance of δ_d. $t_{start} = t_{a,i}$ represents the activity start time, $t_{end} = t_{b,i+1}$ represents the activity end time, and $t_{duration}$ to be the activity duration and is calculated as follows:

$$t_{duration} = t_{b,i+1} - t_{a,i}. \tag{3}$$

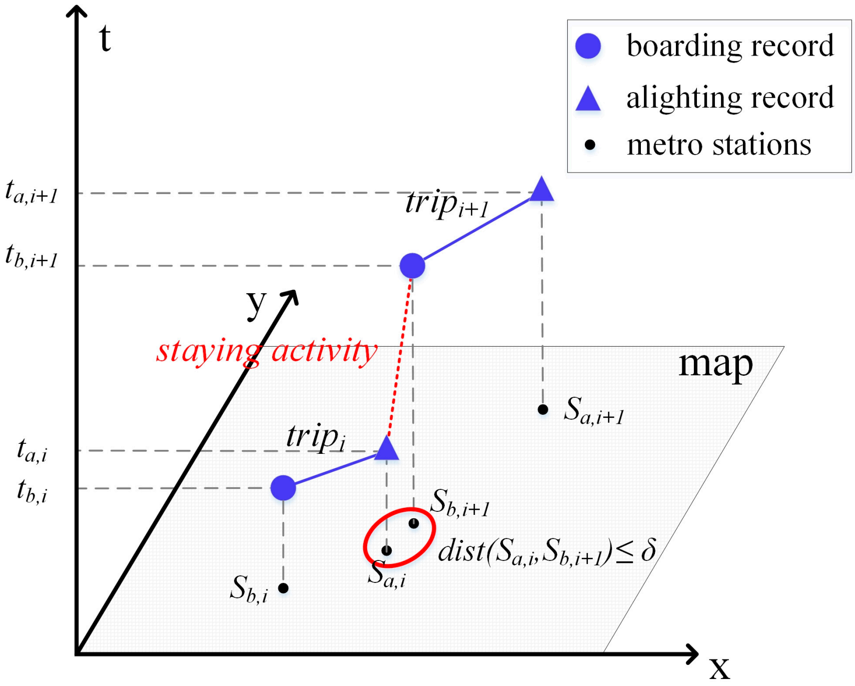


Figure 2. Illustration of staying activity.

4.2. Clustering Staying Activities

The types of activities are different from each other in terms of when the activity starts and the duration that it lasts. An activity that lasts for 8 h is likely to be working, while a short activity around 3 h has a high possibility to be recreation, like shopping or eating. The time when the activity starts also matters. For example, a long stay starting in the morning could be labeled as working, while a long stay that happened in the evening is resting at home. Indeed, it is possible to infer the purpose of each activity according to its start time, duration, and location, along with the help of land use data or travel survey data, as presented in [21]. However, in this study, we pay more attention to varying activity patterns around a metro station in aggregation. Therefore, we cluster activities and label each cluster by its typical temporal feature.

We adopt the start time, the end time, and the frequency as three features to cluster the most typical and frequent activities. We first divide a day into 24 time windows, then introduce a vector X_p to depict the total activities starting in the time window T_p^{start} and ended in the time window T_p^{end}, as denoted in Equation (4):

$$X_p = (T_p^{start}, T_p^{end}, N_p), \tag{4}$$

where N_p is the count of all users' activities that started in T_p^{start} and ended in T_p^{end}. It is standardized by a z-scores method that is used to compare an observation at a dimensionless quantity level; see Equation (5).

$$N_p{}' = \frac{N_p - \rho}{\sigma}, \tag{5}$$

where ρ and σ are the mean and standard deviation of N_p. $N_p{}'$ is positive when the number of activities are above the mean, while $N_p{}'$ is negative when the number of activities are below the mean.

To get the most typical and frequent activities that passengers make, we take the density-based spatial clustering of application with noise (DBSCAN) algorithm to extract the most frequent and typical activities that passengers exhibit. DBSCAN algorithm is a widely used clustering method that does not need to predefine the number of clusters and is robust to noise [41]. There are two key parameters required in the DBSCAN algorithm: the minimum number of points (*MinPts*) and the distance (ε). If there are at least *MinPts* points within a distance ε from point X_p, then X_p is considered to be a core point, which forms a cluster together with all points that lie within the density-reachable distance ε. Otherwise points are marked as outliers if they are neither core points nor border points.

The process of the applied DBSCAN method is as follows:

Step 1: Choose an arbitrary unvisited point, mark it as visited, and retrieve its ε-neighborhood.

Step 2: For the initial point: if there are *MinPts* points in its neighborhood, mark the point as the core point and the current processing point, label it with a new cluster, then go to step 3. Otherwise, go to step 5.

Step 3: For the current processing point: retrieve its neighborhood points within the distance ε into a new seed set. For every point in the seed set, if the point has been marked as an outlier, change its cluster to the current cluster number; if the point is unvisited, go to step 4. If all points in the seed set are visited, go to step 1.

Step 4: Mark the point as visited; label it into the current cluster of the core point; query its neighbors within the distance ε and add them into the current seed set.

Step 5: Mark the point as an outlier and repeat step 1.

Step 6: Stop if all points are visited.

DBSCAN defines clusters as dense regions, separated by regions of lower point density. As a result, we are able to get n activity clusters, denoted as $\mathrm{AC}_1, \ldots, \mathrm{AC}_i, \ldots, \mathrm{AC}_n$. The outliers that have not shown aggregated patterns are labeled as AC_0. To make each cluster more easily interpreted, each type of staying activity is further labeled by the start and duration of the staying activity, such as long or short, daytime or night.

4.3. Inferring Metro Station Area Functions

Using the proportions of passenger staying activities, the function of each metro station area can be inferred. For each metro station s_k, the different patterns of activities happen around the station could be represented as:

$$s_k = [p_{s_k}(\mathrm{AC}_0), \ldots, p_{s_k}(\mathrm{AC}_i), \ldots, p_{s_k}(\mathrm{AC}_n)], \tag{6}$$

where $p_{s_k}(\mathrm{AC}_i)$ is the percentage of activities labeled as AC_i that happen around s_k, and is calculated as:

$$p_{s_k}(\mathrm{AC}_i) = \frac{\mathrm{N}(\mathrm{a}_m | s_{activity} = s_k, \mathrm{a}_m \in \mathrm{AC}_i)}{\mathrm{N}(\mathrm{a}_m | s_{activity} = s_k)}, \tag{7}$$

where $\mathrm{N}(\mathrm{a}_m | s_{activity} = s_k)$ is the total activity count that happened in the anchoring station s_k, and $\mathrm{N}(\mathrm{a}_m | s_{activity} = s_k, \mathrm{a}_m \in \mathrm{AC}_i)$ presents the activity count of AC_i at s_k. As each type of activity has a different base number after the aggregation in Section 4.2., to equally consider the contribution of each type of AC_i, the weight of each AC_i is calculated as:

$$w(\mathrm{AC}_i) = \frac{\frac{\mathrm{N}(\mathrm{a}_m)}{\mathrm{N}(\mathrm{a}_m | \mathrm{a}_m \in \mathrm{AC}_i)}}{\sum_{i=0}^{n} \frac{\mathrm{N}(\mathrm{a}_m)}{\mathrm{N}(\mathrm{a}_m | \mathrm{a}_m \in \mathrm{AC}_i)}}. \tag{8}$$

Therefore, the distance between the vectors of two stations $d(s_{k1}, s_{k2})$ is calculated through the weighted Euclidean distance, as shown in Equation (9):

$$d(s_{k1}, s_{k2}) = \sqrt{\sum_{i=0}^{n} (w(\mathrm{AC}_i) \times [p_{s_{k1}}(\mathrm{AC}_i) - p_{s_{k2}}(\mathrm{AC}_i)])^2}. \tag{9}$$

The activity patterns are highly associated with the land use distribution around the station areas. We aim to extract the stations with similar patterns of staying activities shown by passengers, so we adopt a k-means clustering method to cluster the patterns of activities that happened around the metro stations. Accordingly, we are finally able to set the function of each station area, including residential, commercial, business, and mixed function. The clustering patterns around the station reflect the vibrancy and composition of human activities in the dynamic space. We will discuss the characteristics of each cluster in Section 5.3.

5. Results

5.1. Identification of Staying Activities

According to the staying activity definition presented in Section 4.1, we set $\delta_d = 1$ km, considering the average distance between any two stations is 1034 m in the Wuhan metro system. Furthermore, we set $\delta_t = 24$ h to focus on daily activities. Of all the transaction records during the 23 weekdays in August 2016, we detected about 12.8 million staying activities performed by 2.5 million cardholders. These staying activities take 40.0% of all 32 million travel trips, which means that every 2.5 travel trips will generate on average one staying activity in the dataset. This ratio could also be explained as the ratio of round trips that occurred in the metro system. The average number of staying activities per cardholder is 5.12 per month.

The histogram of staying duration $t_{duration}$ is plotted in Figure 3. There are three obvious peak values from the density curve shown in the plot. The first is the duration that lasts from 1 to 3 h, which are the short activities that people perform. The other peak value is located at about 9–10 h, which has a high probability to be office time, say, from 8:00 a.m. to 6:00 p.m. The third peak value is found at about 13 h, which may imply long work times by hard workers or staying at home. Similar to [38], durations that last 6 or 7 h are very infrequent, which implies that 6–7 h is a natural boundary to distinguish between short and long activities. These distinct patterns show that the duration of a staying activity is a useful index to capture the features of different activities.

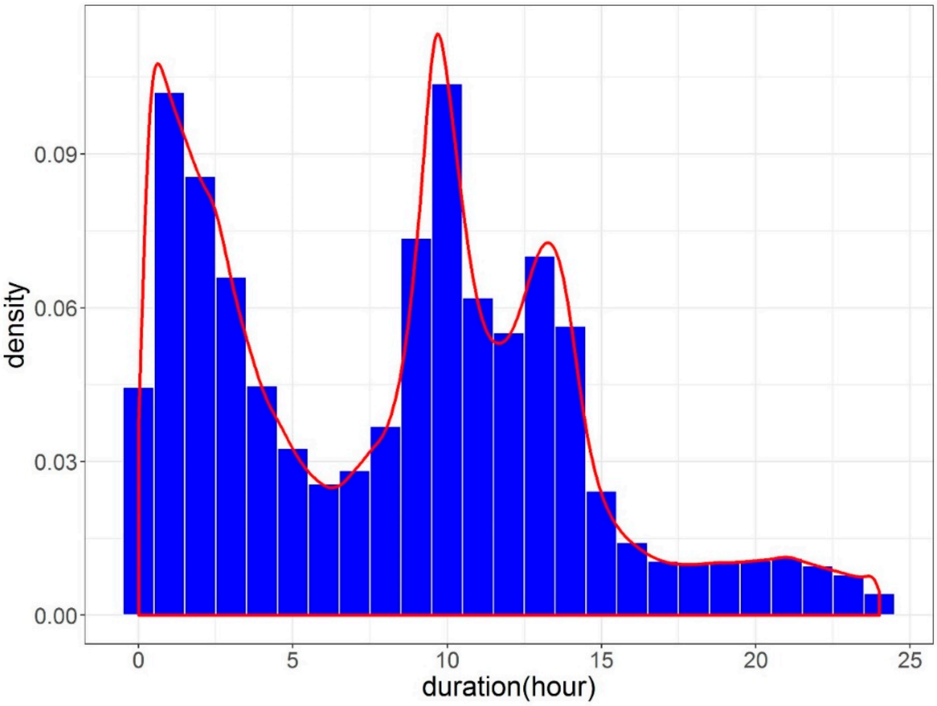


Figure 3. The distribution of staying activity duration (bin width = 1).

The distribution of start time and end time of X_p are plotted in Figure 4. The hours above 24 h in the y axis in the plot denote the time of the next day. The value of each pixel is $N_p{}'$ in Equation (5). We can see that there are several highlighted aggregated clusters, which could be referred to as the peak values in Figure 3. For example, the high density of activities that occur at t7 to t9 and end at t17–t20, a total of 10–13 h, suggests work activity. Another high density occurs at t17–t19 and ends at t31–t32 (7 a.m.–8 a.m. the next day), which is obviously home activity. The short activities happened all day from t8 to t19. Figure 4 gives us more information about when the short or long activities took place in Figure 3.

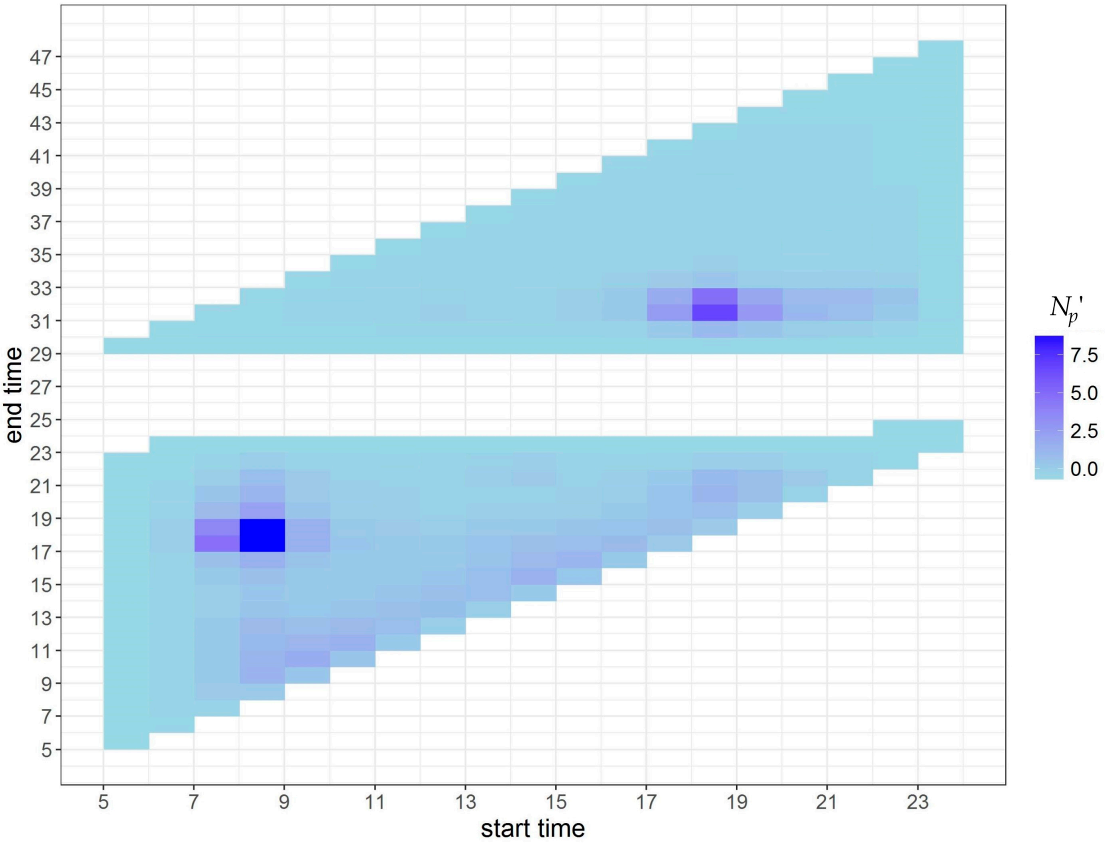
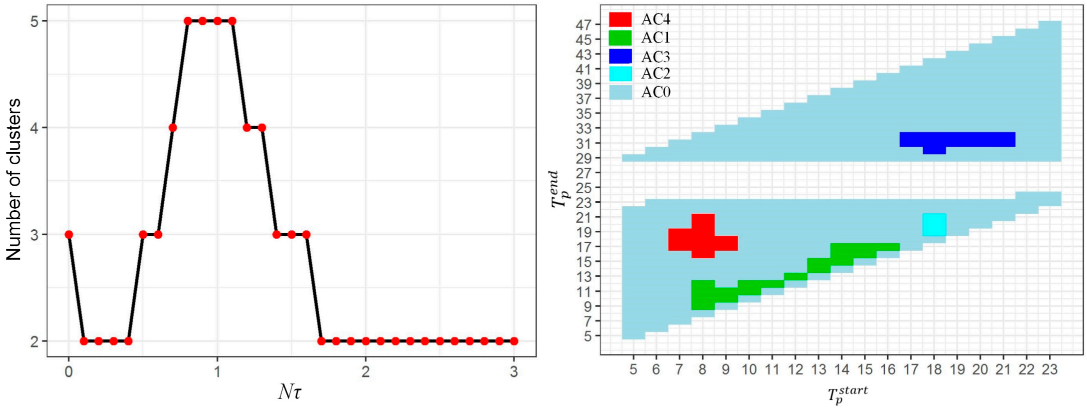


Figure 4. The temporal distribution of start time and end time of X_p.

In the process of DBSCAN, *MinPts* controls the minimum number of observations to form a cluster and is set to be 3; the distance ε is set to be 2 h. In order to focus on clusters that have high intensity, we first select the X_p value that satisfies $N_p{}' \geqslant N_\tau$. Figure 5a shows how the cluster size will change with N_τ. When N_τ is around 0.8~1.1, there would be five clusters as a result, in which the short activities that happened in the morning and the short activities that happened in the afternoon are separate clusters. The correlation between these two activities is examined. The Pearson index is 0.789, which implies that there is a high correlation between them. Herein, it is reasonable that the two activities could be combined as one cluster. Therefore, we chose $N_\tau = 0.7$ and finally were able to get the four typical clusters; the results are shown in Figure 5b. The four clusters are denoted as AC1, AC2, AC3, and AC4, respectively. The rest of the activities that do not belong to any of these clusters in X_p are assigned to AC0. According to their start time and last time in each cluster patterns, we labeled five typical activities: short_daytime (AC1), short_night (AC2), long_daytime (AC4), long_overnight (AC3), and others (AC0). The number of activities in each cluster is listed in Table 2.



(a) (b)

Figure 5. (**a**) How the number of clusters changes with N_τ; (**b**) the cluster results of DBSCAN.

Table 2. Statistics of five cluster activities.

Types of Activities		Count	Cardholder Count	Activities per Person per Month
Short activities	AC1: short_daytime	1,793,508 (14.0%)	969,387	1.85
	AC2: short_night	287,831 (2.3%)	224,005	1.28
	subtotal	2,081,339 (16.3%)	1,104,844	1.88
Long activities	AC3: long_overnight	1,956,964 (15.3%)	434,651	4.50
	AC4: long_daytime	2,644,323 (20.7%)	512,017	5.16
	subtotal	4,601,287 (36.0%)	621,339	7.41
AC0: others		6,099,015 (47.7%)	1,979,297	3.08
total		12,781,641 (100.0%)	2,495,304	5.12

As illustrated in Table 2, of all the staying activities detected, the long activities take 36%, while short activities take 16.3%. The lower count of short activities per person per month (1.88) is much lower than that of the long activities per person (7.41) is partly because the cardholder count (1,104,844) is much larger than that of the long activities (621,339). The average long activities taken per person per month is 7.41, which is higher than average long overnight activities per person (4.50) and average long daytime activities per person (5.16). This implies that there is an overlap between people performing AC3 and AC4. These people are regular commuting groups. The spatial distribution of clusters in each cluster is further discussed in the next section.

5.2. Results of Metro Station Area Functions

We adopt *k*-means method to identify the social function of metro station areas based on the degree of similarity with respect to the proportions of different anchor staying activities. According to Equation (8), the weight in this study is w = (0.03, 0.1, 0.7, 0.1, 0.07). The number of clusters is determined through the Bayesian Information Criterion (BIC), as suggested in [42]. Finally, we get six clusters as a result, and the average value of each cluster is displayed in Figure 6. It is obvious that different clusters have distinct activity patterns. For example, F6 has a highlighted percentage of short activities at night, while F1 typically has long overnight activities, which implies that stations in the two clusters have different social functions. We further map the spatial distribution of six identified clusters as well as examples of each function, as shown in Figure 7.

The function of a metro station area is highly related to the land use distribution of the surrounding area. The average percentage of land use types in each cluster is reported in Table 3. We then compare our results of inferred social functions with the surrounding land use distribution of each metro station in order to get a deeper understanding of the relationship between the dynamic human activity and the static land use pattern. Here we set 500 m as the buffer distance when generating the station catchment area, as suggested (see [43]). We further analyze the spatial distribution of six identified clusters with respect to the land use distribution to gain a better understanding of the human activities at the station level.

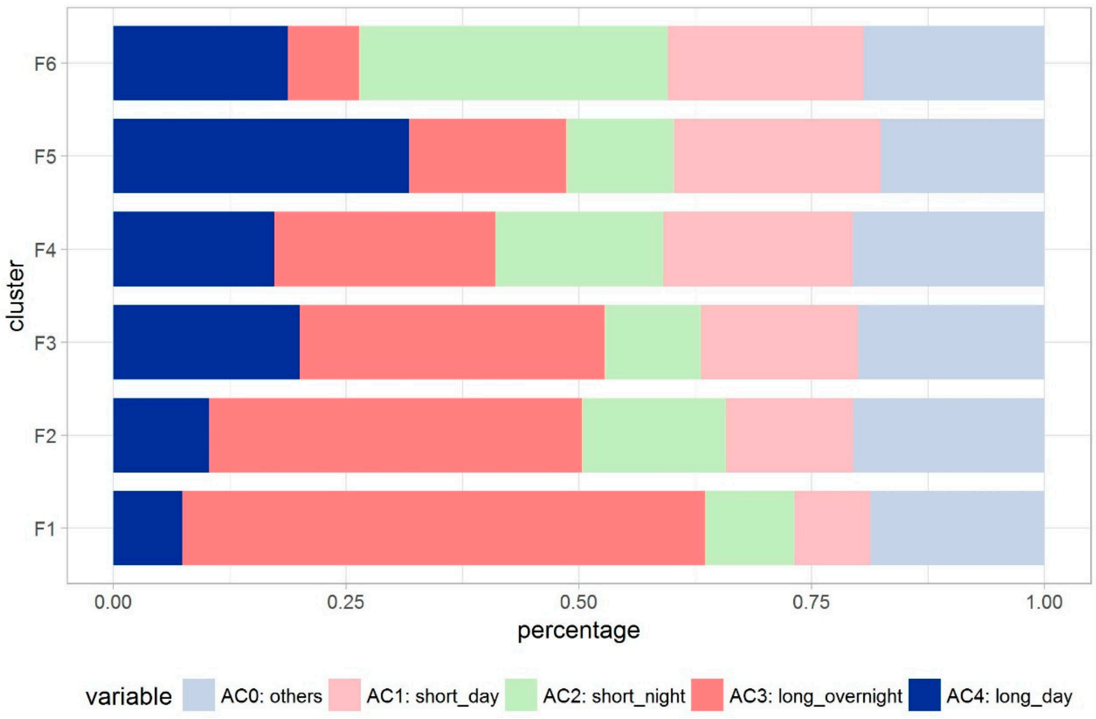


Figure 6. Clustering patterns of different staying activity percentage.

Table 3. The average percentage of land use types in the station area of each function cluster.

Clusters	Adm	PS	Bus	GS	Res	Ind	Tra	LW	MU	Oth
F1	0.8	7.9	2.8	10.0	38.9	2.4	0.1	0.1	1.4	35.6
F2	1.3	9.1	2.9	6.7	45.2	8.1	2.4	2.9	2.3	19.1
F3	2.3	5.7	10.0	5.8	36.1	9.7	3.8	2.4	1.6	22.6
F4	3.8	16.4	10.9	7.3	39.5	2.8	5.8	0.1	2.1	11.3
F5	4.0	10.4	18.1	7.3	32.5	5.0	4.2	0.9	1.4	16.3
F6	5.1	12.7	23.8	6.5	39.1	0.8	1.2	0.1	1.2	9.4

Adm: administrative; PS: public service; Bus: business and commercial; GS: green space; Res: residential; Ind: industrial; Tra: transportation; LW: logistics and warehouse; MU: municipal utilities; Oth: others.

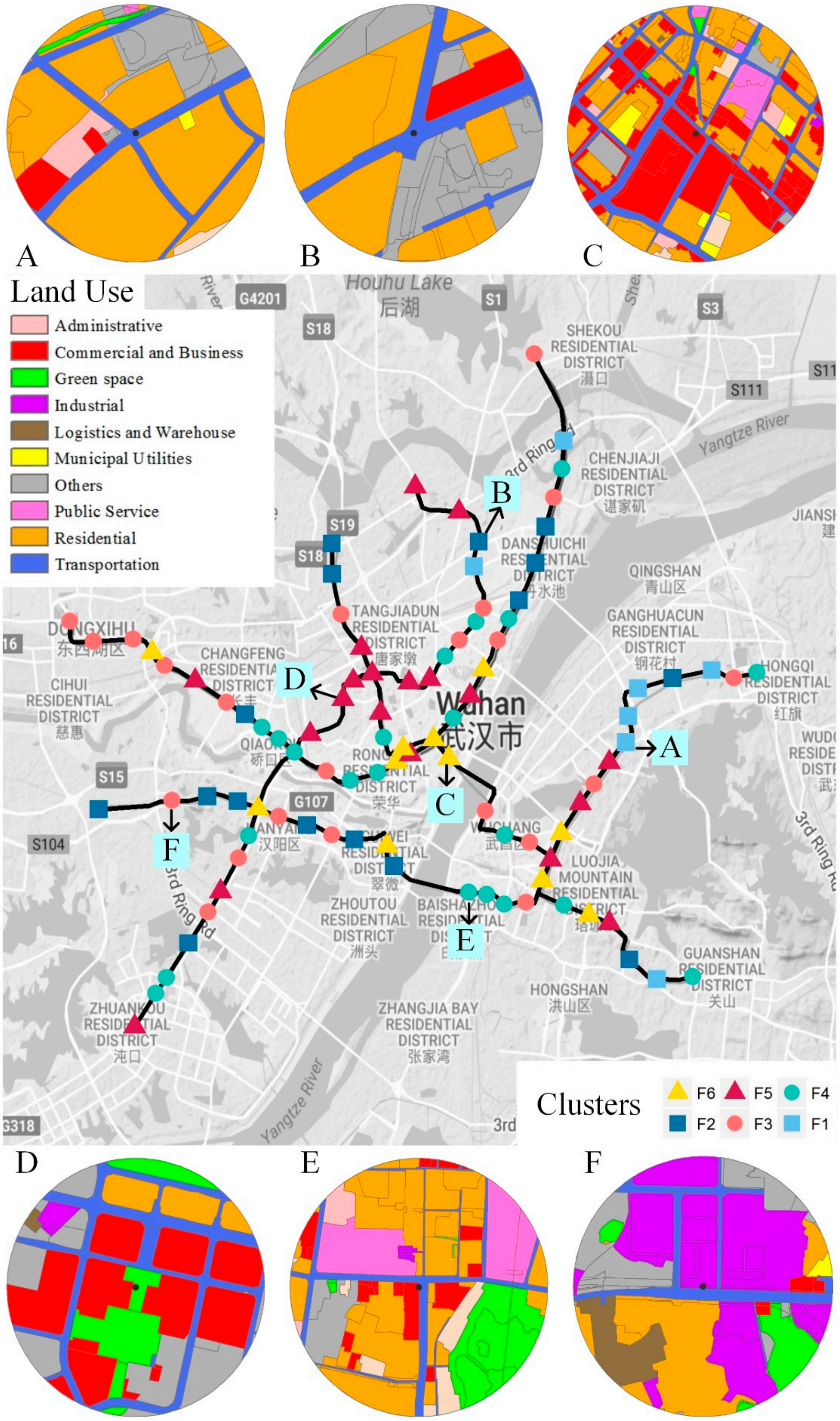

A
B
C
Land Use
Administrative
Commercial and Business
Green space
Industrial
Logistics and Warehouse
Municipal Utilities
Others
Public Service
Residential
Transportation
Houhu Lake
后湖
Wuhan
武汉市
Yangtze River
3rd Ring Rd
SHEKOU RESIDENTIAL DISTRICT
CHENJIAJI RESIDENTIAL DISTRICT
DANSHUICHI RESIDENTIAL DISTRICT
QINGSHAN
青山区
GANGHUACUN RESIDENTIAL DISTRICT
HONGQI RESIDENTIAL DISTRICT
TANGJIADUN RESIDENTIAL DISTRICT
CHANGFENG RESIDENTIAL DISTRICT
DONGXIHU
东西湖区
CIHUI RESIDENTIAL DISTRICT
LUOJIA MOUNTAIN RESIDENTIAL DISTRICT
GUANSHAN RESIDENTIAL DISTRICT
HONGSHAN
洪山区
ZHOUTOU RESIDENTIAL DISTRICT
ZHANGJIA BAY RESIDENTIAL DISTRICT
ZHUANKOU RESIDENTIAL DISTRICT
Clusters
F6
F5
F4
F2
F3
F1
D
E
F

Figure 7. Spatial distributions of six function clusters. A: Tieji Road; B: Houhu Avenue; C: Jianghan Road; D: Wuhan CBD; E: Fuxing Road; F: Mengjiapu. Station A–F are examples of the six clusters respectively.

As mapped in Figure 6, F1 and F2 cluster have the highest percentage (56.1% and 40%) of AC3 long overnight stay among all clusters. People visiting these stations mostly aim to return home.

This gives strong evidence that the function of these two types is typically residential. The difference between the two is that F2 exhibits more short activities in both the daytime (AC1) and the night (AC2), which shows that there are more commercial activities in F2 than in F1. As plotted in Figure 7, station A (Tieji Road) and station B (Houhu Avenue) are examples of F1 and F2. Both station A and station B are basically surrounded by residential land use; station B has larger or more attractive business land use. Although both could be marked with residential function, F2 has a larger mixture of residences with other types of development, like business buildings. Nevertheless, F1 only ranked fourth in residential land use, as shown in Table 3. This shows that the static and dynamic space show some different patterns. We will further discuss this problem in Section 5.3.

As shown in Figure 6, F6 has the highest percentage of AC2 short night activity (33.1%). Short night activity is an index for entertainment or recreation, indicating that the function of these stations is commercial. The land use correspondingly shows that F6 has the highest likelihood of being business type, see Table 3 and Station C (Jianghan Road) presented in Figure 7. Although the residential occupancy in F6 is very high (39.1%), it only contributes to 7.7% of AC3 long overnight activity. The area of residential and business land use around Station C area are nearly equal. However, stations in F6 emphasize entertaining and commercial functions, which attracts people for shopping, eating, or other entertainment.

The F5 cluster has the highest AC4 long daytime activity (31.8%), indicating that these places are the workplaces in the city. It also has the highest percentage of AC1 short daytime activity among all clusters, indicating that people are socializing or doing business in these places. The high occupancy of administrative (10.4%) and public service (18.1%) land use in Table 3 could give the high AC1 an explanation. Station D (Wuhan CBD) in Figure 7 is a good example of this business function.

The F4 cluster shows mixed function, and has relatively even distributions among all activity types. The percentages of AC0 to AC4 are 20.6%, 20.3%, 18.1%, 23.8%, and 17.3%, respectively. Station E (Fuxing Road) in Figure 7 shows similar land use patterns. Surrounded by a park, a hospital, a historic culture square, and several commercial buildings, station E attracts people to do various activities. Likewise, the F3 cluster also shows mixed function. The F3 cluster has more AC3 long overnight activities and fewer AC2 short night activities compared to F4. As shown in Table 3, F4 is mixed with business, public service, green space and transportation land uses, while F3 is mixed due to a high occupancy of business, industrial, logistics and warehouses, and other land uses; see the example of station F (Mengjiapu) in Figure 7.

The analysis presented in this section could enrich our knowledge of how people use the space near metro stations, which is very costly to derive using tradition survey data. Six clusters are drawn to depict the social functions of metro station areas through the distribution patterns of different types of aggregated human activities. From the map in Figure 7, it can be seen that the business function stations (as shown in triangles) formed different characteristics in three districts in Wuhan. The workplaces in Wuchang have a linear shape around the main roads; the commercial and businesses functions in Hankou are more centered and formed a regional agglomeration, while the stations in Hanyang generally have a more scattered distribution of business. It gives us a picture of the urban structure supported by metro systems. Land use distribution decides where passengers travel but, on the other hand, human activities could change the function of the land use. The results of this study provide us with a comprehensive understanding of human activity patterns in metro systems.

5.3. Discussion

Although the social functions of each cluster in Section 5.2 show many agreements in static land use patterns, there are still a lot of differences between the two. The dominant land usage sometimes fails to exhibit the dominant functions. For example, in the Dazhi Road station, as plotted in Figure 8b, the percentage of residential land use is 75.7%, which ranks first among all stations. The business land use around the station only takes 5.7%, while administrative and public service takes 7.5%. However, the staying activities exhibit mixed function, as shown in Figure 8a. This is for two reasons:

firstly, there are a lot of small businesses inside the residential buildings, which leads to multiple usage of a land parcel; secondly, around the station there are two well-known buildings that cover a small floor area but hold many shops for selling computers, mobile phones, and other electronic devices. The area attracts a lot of people visiting this "computer town" in Hankou to shop for electronic devices, which is revealed as the AC1 short day activity and AC4 long day activities in Figure 8a.

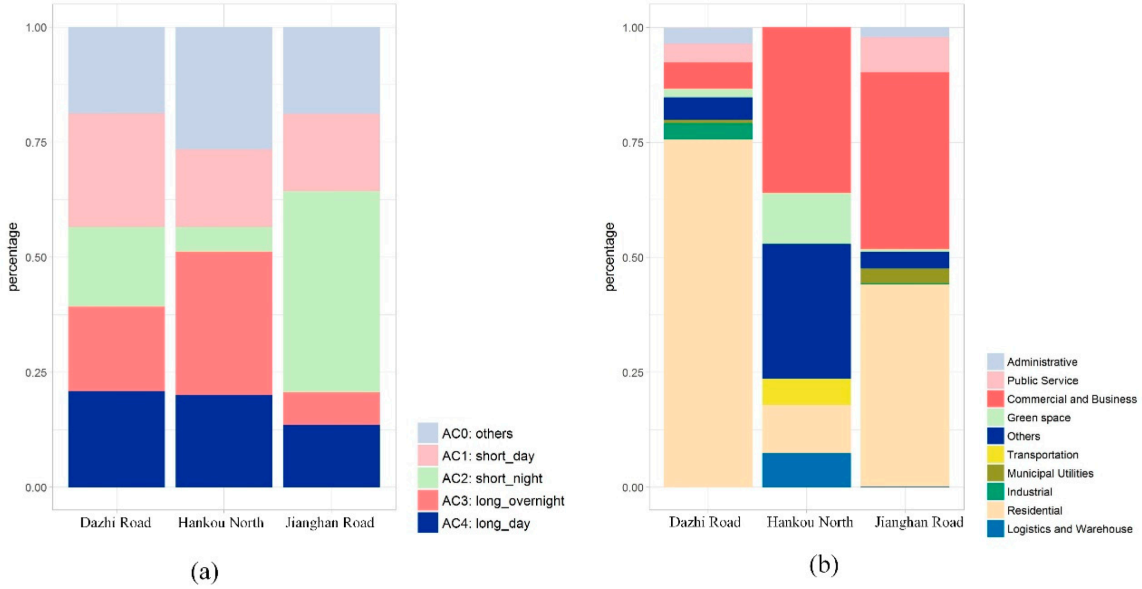


Figure 8. Illustration of three examples: (**a**) the staying activities distribution; (**b**) the land use distribution.

Sometimes, although two stations show similar land usage, the functions are different. Take stations C and D in Figure 7, for example: both stations are surrounded by commercial and business type, but the emphasis of the two is different, thus leading to distinct human activities. Even when two stations are surrounded by commercial buildings, the activity would be different. Take Jianghan Road and Hankou North for example, as depicted in Figure 8b: they both have very high business land usage. However, the activity patterns are quite different (Figure 8a). Jianghan road has a high percentage of short activities, while Hankou North has a noticeable percentage of long overnight staying activities. The stores by Hankou North station are oriented around wholesale trade, while the stores by the Jianghan Road are oriented around retail. The different business models leads to differential attractiveness to visitors, which results in a different distribution in people's anchoring activities attracted by metro transportation. The other factor causing long overnight activities in Hankou North is that the station is located at the end of Line 1. People may transfer to other transportation methods to arrive at their destination after they alight at Hankou North station. So the catchment area of this station could be far greater than 500 m. In other words, using the static land use round the terminal stations of a metro line to make station function inferences may generate errors because the range of the catchment area around the station is hard to determine.

According to the above discussion, we find that there is a gap between the dynamic human activity space and the static land use space around metro stations. The area scale of the land parcel could not be used as a single index to determine how people use space around metro station areas. Even under the same land use type, different human activity patterns are exhibited. Some researchers have realized the same problem and suggested using the distribution of points of interests (POIs) to interpret mixed use within neighborhoods [4,44]. On the other hand, some similar activity patterns may be caused by different land use distribution. This section discusses this differentiation around metro stations; however, the mutual relationships between the two need further investigation on how and why they impact each other.

6. Conclusions and Future Work

The socioeconomic function of the catchment area of a metro station is an important factor when implementing the TOD strategy. However, it is hard to derive the station area's functions using static land use distribution and other traditional survey data. The emergence of smart card data brings us a new opportunity to investigate this problem from the human behavioral perspective. Instead of quantifying the number of check-in and check-out events, we pay more attention to the anchoring activities that happen at the alighting stations. First, we defined the staying activities by the spatial and temporal constraints of the two consecutive alighting and boarding records. Then, by using a DBSCAN method, all staying activities were clustered and labeled via the features of duration, frequency, and start time. By doing this, we were able to get the distribution of aggregated human activities around each metro station. The function of each station area is inferred through the activity distribution characteristics with the help of weighted *k*-means classification. Finally, we analyzed the results of the Wuhan metro systems and discussed the similarities and differentiations between the functions and the land use distribution around the station catchment area.

We got six distinct clusters to depict the functions of all metro station areas in Wuhan. F1 and F2 exhibit a residential function, although F2 has a larger mixture of residential with other types of development than F1. The F5 clusters are typically working places of the city, while the F6 clusters show entertainment and commercial functions. F3 and F4 are both mixed use; however, they are mixed because of different land uses. These findings could give deeper insight into how people act around metro stations. The functions of each station could provide a reference for planners when designing the TOD strategy.

There are several limitations that could be alleviated in future work. Firstly, we use an efficient and easily implemented method to get the labeled feature of activities by using the attributes of start time, duration and frequency. However, if more accuracy is required, we would suggest inferring the purpose of each trip first and then cluster the activities by the trip's purpose so as to get more detailed information about the activity. If a more detailed travel purpose is inferred, it would be very helpful to get refined functions of similar activity patterns in this pattern; for example, the differences between F3 and F4 in the case study could be made distinct. This could be implemented with the help of POI information around the station areas. Secondly, the other limitation is that we only used the activities of weekdays. Activities on weekends are ignored because they have different patterns. This problem would be solved after the trip purpose inference process. Thirdly, using only metro smart card data may cause bias in the results. There are other transportation options that people may choose, like private car, taxi, bus, bicycle, walking, etc. Integrating other types of human movement data, or calibrating with POI data, would be helpful to reduce the data bias problem. Last but not least, we will further study the connections between each metro station to find the interaction patterns between stations with the same or different functions. We believe that these findings will enhance our knowledge about human interaction and deepen our understanding of the urban structure, which will ultimately benefit the city when planning improvements in and around metro stations.

Acknowledgments: This research was supported in part by the National Science Foundation of China (Grant # 41231171). We would like to thank Jack Keel for his English language editing and suggestions. We also thank three anonymous reviewers for their helpful comments.

Author Contributions: This research was mainly conceived and designed by Yang Zhou and Zhixiang Fang. Yang Zhou analyzed the data and performed the experiments. Yang Zhou and Zhixiang Fang wrote the manuscript. Xiongwu Fu provided the dataset. Qingming Zhan and Yaping Huang reviewed the manuscript and provided comments that enhanced the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Calthorpe, P. *The Next American Metropolis: Ecology, Community, and the American Dream*; Princeton Architectural Press: New York, NY, USA, 1993.

2. Lee, S.; Yi, C.; Hong, S.P. Urban structural hierarchy and the relationship between the ridership of the Seoul Metropolitan Subway and the land-use pattern of the station areas. *Cities* **2013**, *35*, 69–77. [CrossRef]
3. Higgins, C.D.; Kanaroglou, P.S. A latent class method for classifying and evaluating the performance of station area transit-oriented development in the Toronto region. *J. Transp. Geogr.* **2016**, *52*, 61–72. [CrossRef]
4. Lyu, G.; Bertolini, L.; Pfeffer, K. Developing a TOD typology for Beijing metro station areas. *J. Transp. Geogr.* **2016**, *55*, 40–50. [CrossRef]
5. Tu, W.; Cao, J.; Yue, Y.; Shaw, S.-L.; Zhou, M.; Wang, Z.; Chang, X.; Xu, Y.; Li, Q. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *Int. J. Geogr. Inf. Sci.* **2017**. [CrossRef]
6. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
7. Zhou, Y.; Fang, Z.; Thill, J.-C.; Li, Q.; Li, Y. Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories. *Comput. Environ. Urban Syst.* **2015**, *52*, 34–47. [CrossRef]
8. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plan.* **2012**, *106*, 73–87. [CrossRef]
9. Yang, X.; Fang, Z.; Xu, Y.; Shaw, S.-L.; Zhao, Z.; Yin, L.; Zhang, T.; Lin, Y. Understanding Spatiotemporal Patterns of Human Convergence and Divergence Using Mobile Phone Location Data. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 177. [CrossRef]
10. Liu, Y.; Sui, Z.; Kang, C.; Gao, Y. Uncovering Patterns of Inter-Urban Trip and Spatial Interaction from Social Media Check-In Data. *PLoS ONE* **2014**, *9*, e86026. [CrossRef] [PubMed]
11. Cui, Z.; Long, Y. Perspectives on Stability and Mobility of Passenger's Travel Behavior through Smart Card Data. *arXiv* **2015**, arXiv:1508.06033.
12. Zhou, W. Catch Me If You Can: Detecting Pickpocket Suspects from Large-Scale Transit Records. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 87–96.
13. Ma, X.; Wu, Y.; Wang, Y.; Chen, F.; Liu, J. Mining smart card data for transit riders' travel patterns. *Transp. Res. Part C* **2013**, *36*, 1–12. [CrossRef]
14. El Mahrsi, M.K.; Côme, E.; Oukhellou, L.; Verleysen, M. Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 712–728. [CrossRef]
15. Zhong, C.; Batty, M.; Manley, E.; Wang, J.; Wang, Z.; Chen, F.; Schmitt, G. Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PLoS ONE* **2016**, *11*, e0149222. [CrossRef] [PubMed]
16. Roth, C.; Kang, S.M.; Batty, M.; Barthe, M. Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows. *PLoS ONE* **2011**, *6*, e15923. [CrossRef] [PubMed]
17. Long, Y.; Thill, J. Combining smart card data and household travel survey to analyze jobs—Housing relationships in Beijing. *Comput. Environ. Urban Syst.* **2015**, *53*, 19–35. [CrossRef]
18. Bagchi, M.; White, P.R. The potential of public transport smart card data. *Transp. Policy* **2005**, *12*, 464–474. [CrossRef]
19. Nassir, N.; Hickman, M.; Ma, Z.L. Activity detection and transfer identification for public transit fare card data. *Transportation* **2015**, *42*, 683–705. [CrossRef]
20. Alsger, A.; Assemi, B.; Mesbah, M.; Ferreira, L. Validating and improving public transport origin-destination estimation algorithm using smart card fare data. *Transp. Res. Part C Emerg. Technol.* **2016**, *68*, 490–506. [CrossRef]
21. Kusakabe, T.; Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transp. Res. Part C Emerg. Technol.* **2014**, *46*, 179–191. [CrossRef]
22. El Mahrsi, M.K.; Côme, E.; Baro, J.; Oukhellou, L. Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. In Proceedings of the 3rd International Workshop on Urban Computing (UrbComp 2014), New York, NY, USA, 24 August 2014.
23. Kieu, L.-M.; Bhaskar, A.; Chung, E. Transit Passenger Segmentation Using Travel Regularity Mined from Smart Card Transactions Data. In Proceedings of the Transportation Research Board 93rd Annual Meeting, Washington, DC, USA, 12–16 January 2014; Volume 7013, pp. 1–17.

24. Zhou, J.; Sipe, N.; Ma, Z.; Mateo-Babiano, D.; Darchen, S. Monitoring transit-served areas with smartcard data: A Brisbane case study. *J. Transp. Geogr.* **2017**. [CrossRef]
25. Ma, X.; Liu, C.; Wen, H.; Wang, Y.; Wu, Y.J. Understanding commuting patterns using transit smart card data. *J. Transp. Geogr.* **2017**, *58*, 135–145. [CrossRef]
26. Long, Y.; Liu, X.; Zhou, J.; Chai, Y. Early birds, night owls, and tireless/recurring itinerants: An exploratory analysis of extreme transit behaviors in Beijing, China. *Habitat Int.* **2016**, *57*, 223–232. [CrossRef]
27. Zhou, J.; Murphy, E.; Long, Y. Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. *J. Transp. Geogr.* **2014**, *41*, 175–183. [CrossRef]
28. Liu, L.; Hou, A.; Ratti, C.; Chen, J. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. In Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, USA, 4–7 October 2009; pp. 842–847.
29. Lathia, N.; Quercia, D.; Crowcroft, J. The Hidden Image of the City: Sensing Community Well-Being from Urban Mobility. In Proceedings of the 10th International Conference on Pervasive Computing, Newcastle, UK, 18–22 June 2012; pp. 91–98.
30. 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]
31. Long, Y.; Han, H.; Tu, Y.; Shu, X. Evaluating the effectiveness of urban growth boundaries using human mobility and activity records. *Cities* **2015**, *46*, 76–84. [CrossRef]
32. 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]
33. 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]
34. Zhong, C.; Huang, X.; Müller, S.; Schmitt, G.; Batty, M. Inferring building functions from a probabilistic model using public transportation data. *Comput. Environ. Urban Syst.* **2014**, *48*, 124–137. [CrossRef]
35. 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]
36. Gong, Y.; Lin, Y.; Duan, Z. Exploring the spatiotemporal structure of dynamic urban space using metro smart card records. *Comput. Environ. Urban Syst.* **2017**, *64*, 169–183. [CrossRef]
37. Chakirov, A.; Erath, A. Activity identification and primary location modelling based on smart card payment data for public transport Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport. In Proceedings of the 13th International Conference on Travel Behaviour Research(IATBR), Toronto, ON, Canada, 15–20 July 2012; pp. 1–22.
38. Bouman, P.; Kroon, L.; Vervest, P. Detecting Activity Patterns from Smart Card Data. In Proceedings of the 25th Benelux Conference on Artificial Intelligence, Delft, The Netherlands, 7–8 November 2013.
39. Peng, Y.; Chen, W. Metro-city Planning Practice: Wuhan Example. *Planners* **2016**, *32*, 5–10. (In Chinese)
40. Wuhan Transportation Development Strategy Institute. *Wuhan Transportation Development Annual Report*; Wuhan Transportation Development Strategy Institute: Wuhan, China, 2017. (In Chinese)
41. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 226–231.
42. Pelleg, D.; Moore, A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), San Francisco, CA, USA, 29 June–2 July 2000.
43. Jun, M.J.; Choi, K.; Jeong, J.E.; Kwon, K.H.; Kim, H.J. Land use characteristics of subway catchment areas and their influence on subway ridership in Seoul. *J. Transp. Geogr.* **2015**, *48*, 30–40. [CrossRef]
44. Yue, Y.; Zhuang, Y.; Yeh, A.G.O.; Xie, J.-Y.; Ma, C.-L.; Li, Q.-Q. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. *Int. J. Geogr. Inf. Sci.* **2016**, *31*, 658–675. [CrossRef]