Beijing City Lab

Gao S, Long Y, 2015, Finding Public Transportation Community Structure based on Large-Scale Smart Card Records in Beijing. Beijing City Lab. Working paper #59

Finding Public Transportation Community Structure based on Large-Scale Smart Card Records in Beijing

Song Gao[a], Ying Long[b]
a. University of California, Santa Barbara, CA 93106, USA, Email: sgao@geog.ucsb.edu
b. Beijing Institute of City Planning, Beijing 100045, China, Email: longying1980@gmail.com

Abstract:
Public transportation in big cities is a crucial part of urban transportation infrastructures. Exploring the spatiotemporal patterns of public trips can help us to understand dynamic transportation patterns and the complex urban systems thus supporting better urban planning and design. The availability of large-scale smart card data (SCD) offers new opportunities to study intra-urban structure and spatial interaction dynamics. In this research, we applied the novel community detection methods from the study of complex networks to examine the dynamic spatial interaction structures of public transportation communities in the Beijing Metropolitan Area. It can help to find the ground-truth community structure of strongly connected traffic analysis zones by public transportation, which may yield insights for urban planners on land use patterns or for transportation engineers on traffic congestion. We also found that the daily community detection results using SCD are different from that using household travel surveys. The SCD results match better with the planned urban area boundary, which means that the actual operation data of publication transportation might be a good source to validate the urban planning and development.
Keywords: public transportation, smart card records, spatial interaction, OD flow matrix, community detection, urban big data

1. Introduction

Public transportation in big cities is a crucial part of urban transportation infrastructures. Exploring the spatiotemporal patterns of public trips can help us to understand dynamic human movements, transportation patterns and the complex urban systems thus supporting better urban planning and design. The availability of large-scale smart card data (SCD), which is one type of urban Big Data collected from public transportation operations and management institutions, offers new opportunities to study the intra-urban structure and spatial interaction dynamics by zooming into individual-based public trips. Previous research has investigated the jobs-housing relationships and commuting patterns using such data and demonstrated comparisons with traditional high-cost travel survey approach (Long et al. 2012, Long & Thill 2013). The study of spatial interactions is one of the traditional researches in geography and regional science. For regional studies, the functional region is defined by regional geographers based on interactions between its distinctive land-use zones (Johnston et al. 1981). Representation forms of spatial interactions between different zones include human movement, commodity flow, resource allocation, information communication and so on. For the past several decades, studies of spatial interaction processes have mainly been based on the census datasets (Rae 2009; Jang and Yao 2011). Recent fast development in information and

communications technology (ICT) and the availability of big geospatial data (such as mobile phone records, GPS-enabled taxi/cab traces, location-based check-ins) has supported several frontier researches on spatial interactions and networks (Ratti et al. 2010, Gao et al. 2013, Kang et al. 2013, Liu et al. 2014), identifying functional urban regions (Manley 2014), as well as to reveal spatiotemporal intra-urban land use variations from travel patterns (Liu et al. 2012).

In this research, we are interested in extracting origin-destination (OD) flow matrices in the aggregation scale of traffic analysis zones (TAZs) and analyzing the intra-urban spatial interaction patterns revealed by human movements among TAZs using public transportation. Traditional spatial clustering approaches, which group similar spatial objects into classes, are not sufficient to explore the network structure of spatial interactions between different regions. Thus we applied the novel community detection methods from the study of complex networks to examine the dynamic spatial structures of public transportation communities in the Beijing Metropolitan Area (16,410 km^2). It can help to find the ground-truth community structure of strongly connected TAZs by public transportation, which may yield insights for urban planners on land use patterns or for transportation engineers on traffic congestion.

This chapter is organized as follows. Section 2 and 3 describe the data and methodology used for this study. We elaborate the detailed results in Section 4. We conclude this work and propose next-step plans in Section 5.

2. Data

In Beijing, most bus/metro passengers use smart cards when getting on and off buses and metros to pay their fares. Thus, individual OD trips which connect bus stops (or metro stations) can be extracted directly from the detailed records of SCD. The collected SCD consists of 97.9 million trips from anonymized 10.9 million smart card users during a one-week period from April 5 to April 11, 2010. More details on public transportation and SCD in Beijing are available in the chapter entitled “Profiling underprivileged residents with mid-term public transit smartcard data of Beijing” in this book. In order to create the public transportation OD flow matrices in the TAZ level, we first georeferenced all bus stops and metros stations with latitude/longitude coordinates, and then spatially joined them into the total 1911 Beijing TAZ boundaries (see Figure 1). A directed-weighted linkage between two TAZs represents the total number of public trips from the origin-TAZ to the destination-TAZ in a given time interval. Regarding the temporal dynamics, we aggregated the data into different hourly and daily periods to study the spatiotemporal patterns in public transportation, as well as variations between weekdays and weekends.

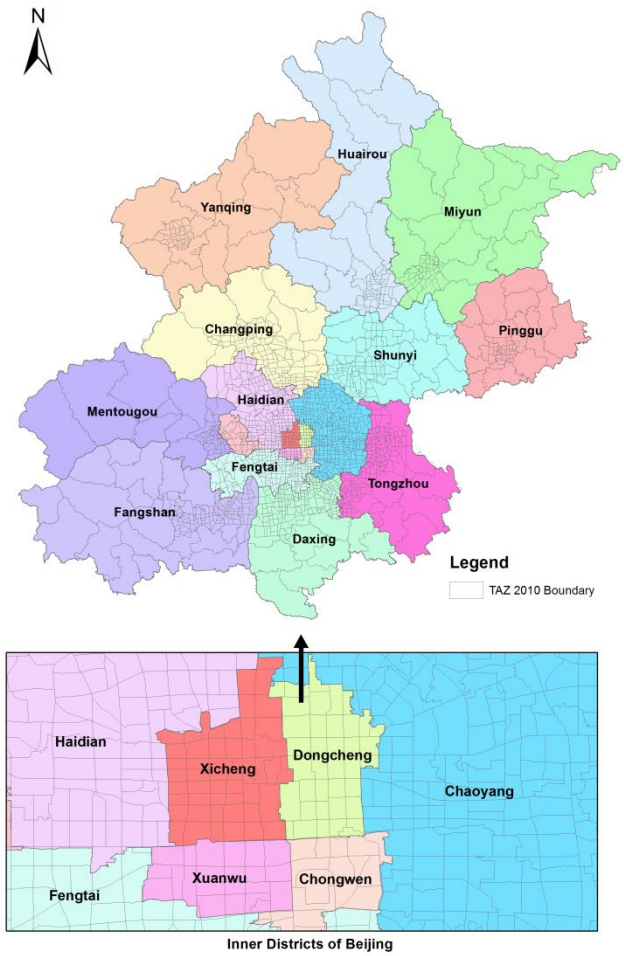


Figure 1. The study area in administrative districts (different colors) of Beijing. It is noteworthy that the basic spatial unit is a TAZ, and the district divisions showed here was the 2010 version without the merges of Xuanwu and Chongwen districts in order to keep consistent with the SCD data collection period.

3. Methodology

In the study of complex networks, a community is defined as a subset (group) of the whole network and the nodes in the same community are densely connected internally and grouped together. The identification of such densely connected nodes in networks is called *community detection*. Popular community detection methods can be classified into two groups: *graph partitioning* and *hierarchical clustering*. Graph partitioning divides a network graph into a set of non-overlapping groups, while hierarchical clustering seeks to build a hierarchy of clusters of nodes, such that for each cluster there are more internal than external connections.

Newman and Girvan (2004) propose a modularity metric to evaluate the quality of a particular division of a network into communities. Modularity compares a proposed division to a null model in which connections between nodes are random. It is defined as the sum of

differences between the fraction of edges falling within communities and the expected value of the same quantity under the random null model.

$$Q = \sum_{k} \sum_{ij \in C} (realflow_{ijk} - estflow_{ijk}) \quad (2)$$

where k is the number of partition communities, $realflow_{ijk}$ gives the actual fraction of interactions between nodes i and j within the same community C, and $estflow_{ijk}$ represents the expected values under the random null model or other theoretical models. If the fraction of edges within communities is no better than the null model the modularity Q=0, while Q=1 indicates the most robust community structure. In practice, modularity values of different real world networks with varying sizes fall into the range 0.3 to 0.7.

We first converted the TAZ-scale OD flow matrices in the consecutive seven days into seven undirected-weighted graphs, where each TAZ can be taken as a node and each OD-flow interaction as a weight edge linking two TAZs. Then, the widely used Newman-modularity-maximization method (Newman 2004) was applied to find the daily public transportation communities. In practice, a bottom-up fast greedy algorithm (Clauset et al. 2004, Gao et al. 2013) was adopted for searching an optimized graph partition that maximizes the modularity measure. First, each TAZ started in its own independent cluster of community and the modularity values among all pairs of TAZs for all communities were calculated. Second, a pair of TAZs which has the maximum difference of OD flow compared with the null model should be merged into a community. Third, the modularity of the new graph will be calculated again and then repeating the procedure until the maximum of modularity is found. A larger modularity value indicates a more robust community structure.

In the following section, we conduct different community detection experiments on SCD in different temporal scales and will further explain the identified community structures in detail.

4. Results

4. 1 Identification of Communities on Weekdays and Weekends

We first examine the daily publication transportation community structure. Table 1 shows the detailed network information of daily community detection results of public transportation OD trips during a week. We find the community consistent pattern in terms of the number of divided groups (6), the average size of each community (202) and the maximum value of modularity (0.457~0.475) in the detection processes. The community on Thursday has the largest modularity value, which indicates a more stable network community structure than other days.

Table 1 Daily community detection results of public transportation OD trips in a week

Day of Week	# of Nodes	# of Edges	# of Groups	Mean of Community Size	MAX Q
Monday	1214	55205	6	202	0.467
Tuesday	1214	55598	6	202	0.461
Wednesday	1216	55510	6	203	0.462
Thursday	1213	55641	6	202	0.475
Friday	1213	55614	6	202	0.457

Saturday	1212	55490	6	202	0.470
Sunday	1213	53273	6	202	0.471

Although the network statistics are similar in the seven days of a week, the spatial distributions of these detected communities lie in slightly different. As shown in Figure 2, the daily community detection results demonstrate that in general geographically cohesive regions that correspond well with administrative districts (such as *Tongzhou*) or merged boundaries (such as *Fengtai* and *Daxin*) in Beijing were identified by weekday public transportation patterns, while some unexpected spatial structures might uncover hidden urban structure that needs further investigation. The suburb public transportation communities usually contain more TAZs than urban central TAZs. There exist strong public transit connections among TAZs which locate along the middle west-to-east corridor including the *Chang'an Avenue* in Beijing, where the Beijing Subway line 1 also runs through the street. Note that the passengers can use smart card when they travelled on metro lines. Surprisingly, most of the southern TAZs in *Daxing*, *Fengtai* and *Fangshan* districts were aggregated into a large transportation community. It indicates there are more frequent intra-public trips within its own community in the southern region than the inter-community trips across other sub-regions of the Beijing Metropolitan Area. The same giant community pattern lies in the northwest TAZs in *Yanqing* district and the majority of TAZs in *Tongzhou*, although there is several connected TAZs from inner districts to *Tongzhou* through the *Beijing Subway Batong line*. Also, it is remarkable to see an enclave in the southern part of *Fangshan* district has been aggregated into a spatially separated large community north/northeast parts of Beijing (covering a large portion of *Chaoyang*, *Shunyi* and *Changping* districts) in all seven consecutive days, which indicates a strong public transportation connection pattern. The integrated analysis of geographical contexts, land-use types, housing prices, job opportunities, and the prominent points of interest in these regions might offer better explanation about the patterns identified in the community detection results. In addition, a northwest TAZ in *Changping* district was aggregated into a large number of spatially separated TAZs which belong to inner districts of Beijing (*Dongchen, Xicheng*, *Chongwen* and *Xuanwu*) only in weekends not in weekdays. It reveals a recreation place of interests in the northwest TAZ and attracts a large portion of public travel trips. Potentially, this pattern could help local transportation agency to identify the needs to provide temporal services for increasing public transportation demands in these connected regions.

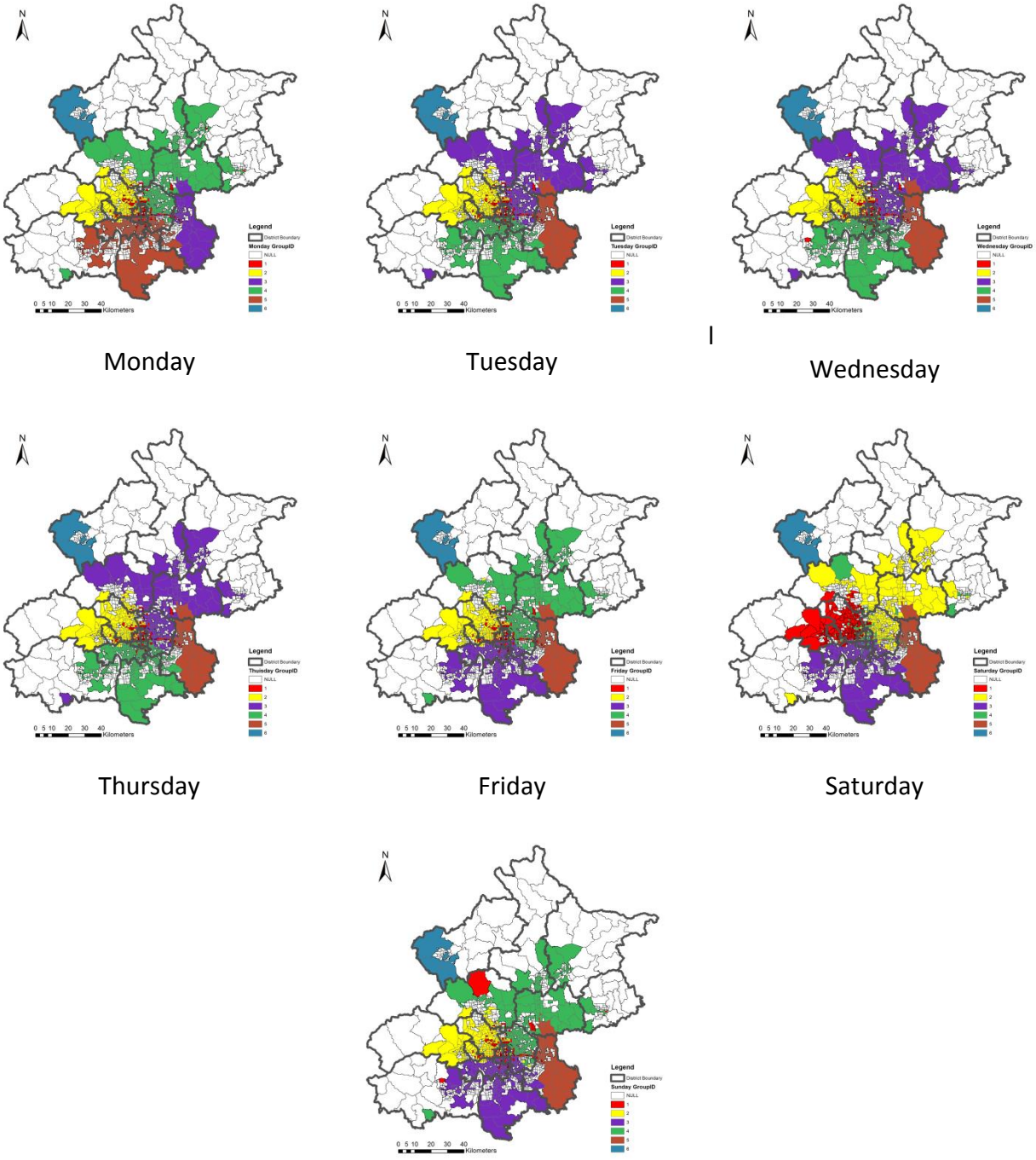


Figure 2. The spatial distributions of daily community detection results of public transportation OD trips using SCD in a week

We also created an interactive web map for exploring the public transportation community detection results in the geographical context (Figure 3). The online geovisualization of communities for comments and validation using local knowledge can be accessed at

http://www.beijingcitylab.com/projects-1/3-bus-landscapes/. We have invited online browsers to propose comments on the communities detected.

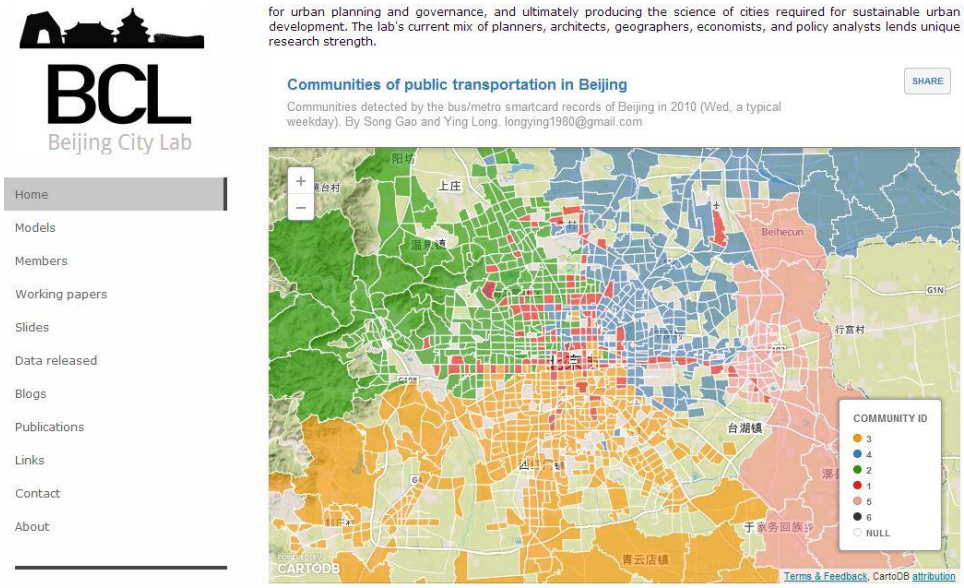


Figure 3. The interactive web map for exploring the public transportation communities in Beijing

4. 2 Comparison with Household Survey Data

Household travel survey is a traditional data-collection approach for acquiring information about residents' travel behaviors and estimating transportation demands (Beijing Transportation Research Center, 2011). The survey tracks travelers' socio-economic attributes, as well as trip origin and destination, time and duration, purpose and mode. We used the survey in 2010 for comparing with results from SCD. The sample size of the survey is 46,900 households (116,142 residents) in the whole Beijing Metropolitan Area. The 2010 Survey provides the one-day travel diary of all respondents covering all travel modes. We applied the same data processing and community detection procedure introduced above to one-day household survey data. We found that the daily community detection results using the household travel surveys are different from that using SCD.

As shown in Figure 4, there are thirteen communities identified when maximizing the modularity of the TAZ network connected by survey OD trips. It is clear to see that most of the TAZs in suburbs have been grouped into the corresponding outer districts. The community boundaries generally match well with administrative boundaries of Beijing Districts. For those places that didn't match, especially for the spatial separated communities, it usually indicates some interesting travel patterns, land-use or urban structure, which could be identified through geographical contexts analysis (Gao et al. 2013).

By comparing the community detection results of SCD and household survey data, we also find that the SCD results match better with the Beijing planned urban area boundary (see Figure 5), which means that the actual operation data of publication transportation might be a good source to validate the urban planning and development.

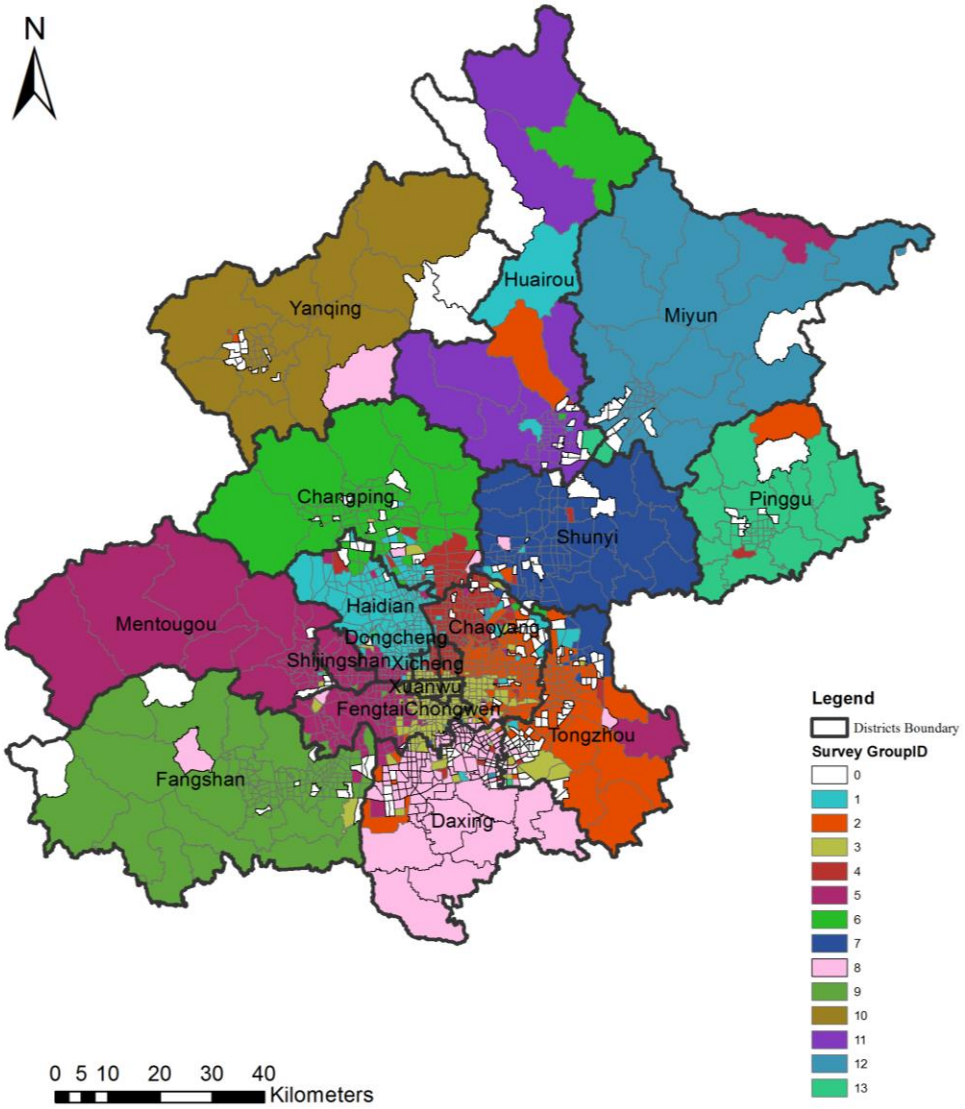


Figure 4. The spatial distribution of community detection results of one-day household survey data

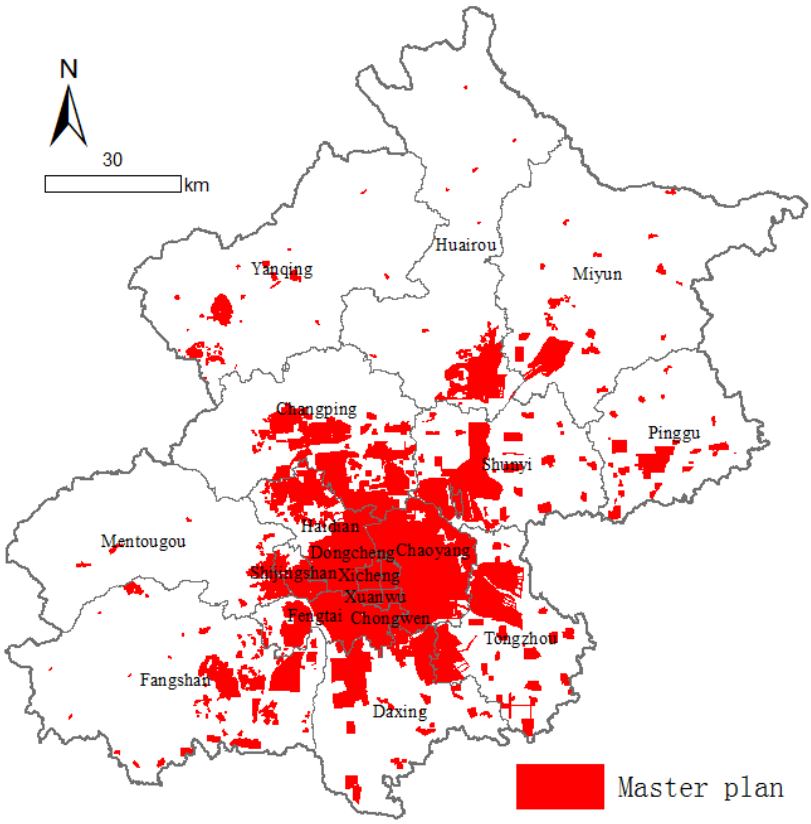


Figure 5. Planned urban area in Beijing

4. 3 Hourly Patterns

Coming back to the SCD, it is also valuable to study the spatiotemporal patterns in a micro-time scale. The community detection results of three-hourly aggregated trips, especially the commuting trips at peak hours yield insights on the overall job-related mobility patterns and intra-TAZ spatial interactions using public transportation. Table 2 shows the detailed network information of three-hourly community detection results of SCD during a weekday. We find that the hourly network structures change more (nodes and edges) than those of daily networks. The maximum modularity in hour 18-21 (0.473) and 09-12 (0.470) has the largest values and thus indicates a more robust community structure. For the spatial distribution patterns (see Figure 6), the northern TAZs change more frequently than the southern parts, especially in the *Changping* District. In addition, similar to the daily patterns, there exist strong public transit connections through the whole day in TAZs that are located along the central west-to-east corridor including the *Chang'an Avenue* in Beijing, where the Beijing Subway line 1 (west-east) run through, as well as those TAZs along the northern part of the subway line 5 (north-south corridor). The interactive web map could also help us to identity underlying patterns by overlaying the detection results on the geographical contexts.

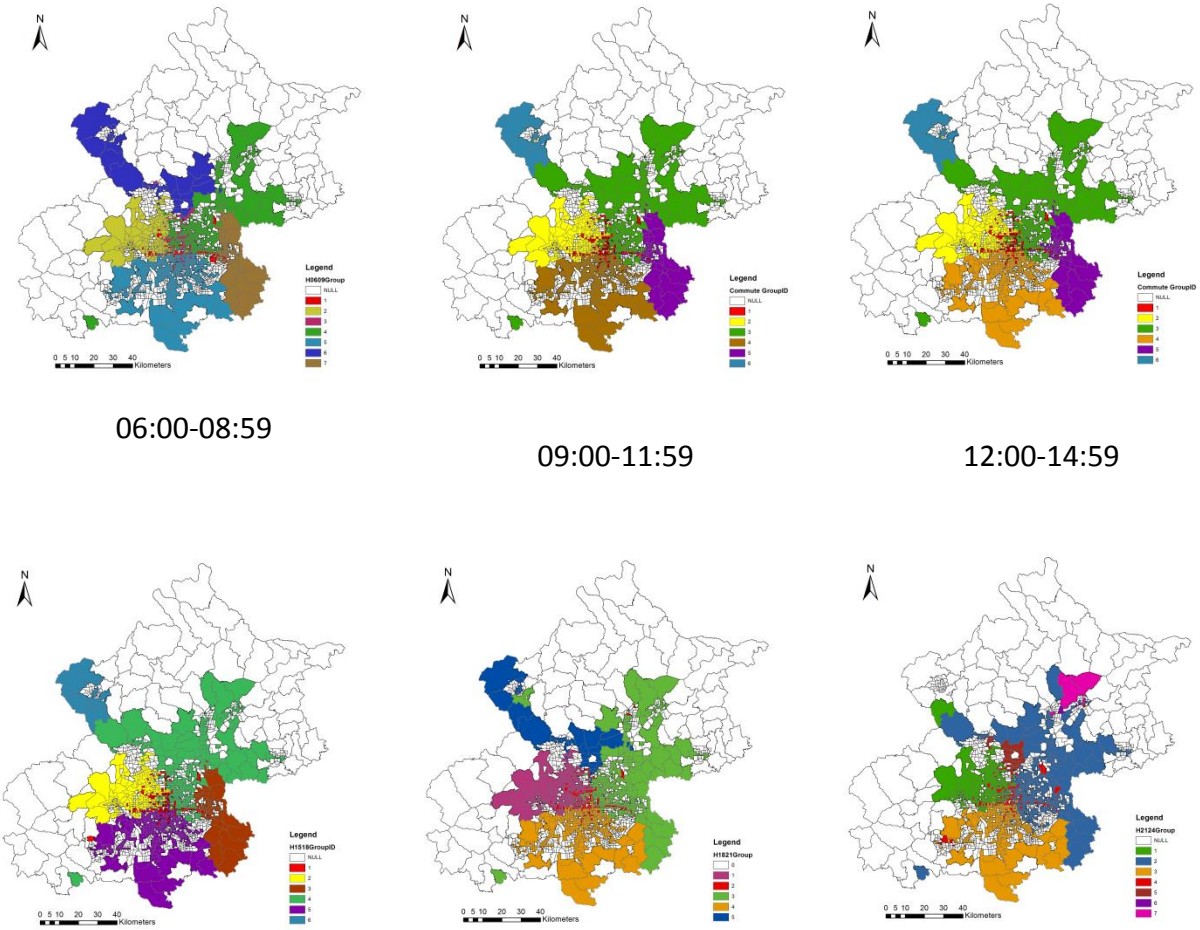


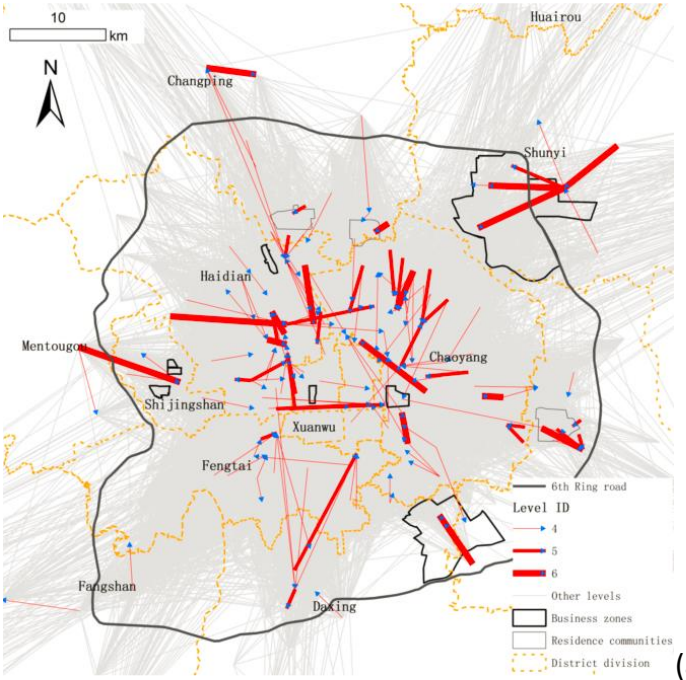
Figure 6. The spatial distributions of three-hourly community detection results of public transportation OD trips using SCD in a weekday

Table 2 Three-hourly community detection results of public transportation OD trips using SCD in a weekday

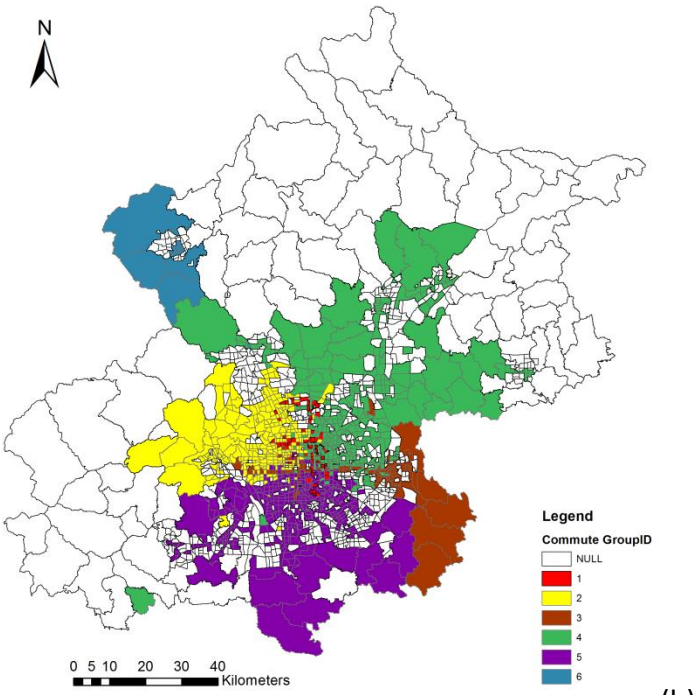
Hours	# of Nodes	# of Edges	# of Groups	Mean of Community Size	MAX Q
06-09	1205	45359	7	172	0.459
09-12	1207	45607	6	201	0.470
12-15	1206	55510	6	203	0.462
15-18	1212	45519	6	202	0.452
18-21	1208	46378	5	242	0.473
21-24	1145	31064	7	164	0.451

4. 4 Identification of Community Structure on Commuting Trips

In big cities, commuting trips usually contribute the largest share of daily public transportation. The commuting behaviors rely on the spatial distribution of regional job opportunities and housing markets. Long et al. (2012) introduced an algorithm to identify the residential zones and job places of the smart card users and further extract their commuting trips in Beijing. We extracted more than 700, 000 commuting trips based on the 2010 SCD using this algorithm. Figure 7(a) shows the prominent home-to-job flows that represent heavy commuting traffic by a head/tail division. The spatial distribution of identified six commuting network communities is shown in Figure 7(b). The largest commuting community contains 375 TAZs, while the smallest one has only 15 TAZs and the mean size is about 197. It tends to have more heavy commuting trips within the same community zones compared with inter-communities. Interestingly, the commuting community result is similar to the hourly patterns between 9AM-12PM instead of that of 6AM-9AM (in Figure 6), but a small modularity value 0.349 indicates that it is not a robust community structure and might vary over different time periods. The formation of commuting communities results from the influence of where people live and work across different zones of the city.



(a)



(b)

Figure 7. (a) The arrow-links in the TAZ scale illustrates the commuting patterns (from home-region to job-region) in Beijing; (b) The spatial distributions of community detection results on commuting trips.

5 Conclusions and Future Work

In this chapter, we applied the community detection methods based on the study of complex networks to examine the dynamic spatial interaction structures of public transportation communities in the Beijing Metropolitan Area using SCD. A community represents that there is a subset of TAZs in which the smart card holders have more intra-community trips than inter-community trips via public transportation (i.e., bus and subway/metro). It also reflects the spatial heterogeneity of OD trips. There are several findings based on our experiment results:

First, the community detection results help to identify the functional connected traffic analysis zones by public transportation and most of them are consistent in both weekdays and weekends. Some detected spatially separated TAZs which belong to the same community indicate strong public travel demands in these regions, either for commuting trips on weekdays or for recreational trips on weekends.

Second, the daily community detection results using SCD are different from that using household travel surveys and the SCD community boundaries match better with Beijing urban planned area than the household travel survey.

Third, the hourly network structures change more than those of daily networks; the community detection results also have more variances in spatial distribution.

Fourth, the identified community structure on commuting trips shields insights on where and which residential- and job-related TAZs are connected by public transportation.

This research applies a network-analysis approach to investigate the ground-truth community structure of strongly connected TAZs via public transportation, which yields insights on urban structure in Beijing from the public transportation functional zone perspective. In further research, we would like to conduct more detailed analysis by integrating land-use data, points-of-interest (POI) database with human activities from household surveys or social media to give a more holistic view of public transportation using emerging urban big data and computing techniques. In addition, the map matching of these OD trips to actual streets and further analysis could be beneficial for the reliability analysis of street networks and emergency transportation management.

References

1. Beijing Transportation Research Center. (2011). *Beijing Transportation Annual Report 2011* (In Chinese).
2. Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111.
3. Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, *17*(3), 463-481.
4. Jang, W., & Yao, X. (2011). Interpolating spatial interaction data.*Transactions in GIS*, *15*(4), 541-555.
5. Johnston, R., Gregory, D., & Smith, D (1981). The Dictionary of Human Geography. Oxford, Blackwell Reference.
6. Kang, C., Zhang, Y., Ma, X., & Liu, Y. (2013). Inferring properties and revealing geographical impacts of intercity mobile communication network of China using a subnet data set. *International Journal of Geographical Information Science*, *27*(3), 431-448.
7. Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai.*Landscape and Urban Planning*, *106*(1), 73-87.
8. Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS one*, *9*(1), e86026.
9. Long, Y., Zhang, Y., & Cui, C. Y. (2012). Identifying commuting pattern of Beijing using bus smart card data. *Acta Geographica Sinica*, *67*(10), 1339-1352.
10. Long, Y., & Thill, J. C. (2013). Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. *arXiv preprint arXiv:1309.5993*.
11. Manley, E. (2014). Identifying functional urban regions within traffic flow.*Regional Studies, Regional Science*, *1*(1), 40-42.

12. Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, *69*(6), 066133.
13. Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113.
14. Rae, A. (2009). From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census.*Computers, Environment and Urban Systems*, *33*(3), 161-178.
15. Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., & Strogatz, S.H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLoS one* 5: e14248