Journal of Transport Geography 66 (2018) 213–223

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Exploring the difference between ridership patterns of subway and taxi: Case study in Seoul

Kyoungok Kim

Information Technology Management Programme, International Fusion School, Seoul National University of Science & Technology (SeoulTech), 232 Gongreungno, Nowon-gu, Seoul 139-743, Republic of Korea

ARTICLE INFO

Keywords:
Taxi trip
Subway trip
Ridership pattern
Urban transit
Classification
Clustering

ABSTRACT

Understanding urban mobility patterns and their connections with different area characteristics is a traditional topic in urban studies, considering its importance for the planning and management of urban facilities, transportation systems, and services. Data recordings about trips using different means of transportation, such as a subway, bus, and taxi have been collected because of the development of IT technologies; such development has motivated various research related to uncovering detailed urban mobility patterns and factors that affect mobility. However, many works usually focus only on a specific means of transportation and fail to present different aspects of ridership patterns for other means of transportation. In this study, subway and taxi data were analyzed simultaneously to uncover factors on human mobility depending on the means of transportation in Seoul. The present research focused on regions nearby subway stations. Data mining techniques, such as clustering and classification, were employed. Different distinct ridership patterns of subway and taxi were detected using clustering; moreover, the difference between ridership patterns and spatial distributions of clusters were examined. A two-step classification analysis was then performed to determine factors that influence ridership patterns.

1. Introduction

Understanding urban mobility patterns and their connections with area characteristics, including land use, demographic information, and existing facilities, is a traditional topic for urban studies because it is important for the planning and management of urban facilities, transportation systems, and services, such as analyzing turnover rates, and trip rates per user (Goodchild et al., 1993; Handy, 1996; Badoe and Miller, 2000; Boarnet and Crane, 2001; Waddell, 2002; Blythe, 2004; Park et al., 2008). Previous data collection of individual trajectories for understanding human mobility primarily depended on surveys, which provided limited information (Hamilton and Röell, 1982; Gordon et al., 1989; Vilhelmson, 1999). However, the recent development of IT technologies has pervaded in cities, which has resulted in the possibility of tracking the daily mobility of inhabitants using massively collected records from different devices such as mobile phones (Ratti et al., 2006; Reades et al., 2007; Gonzalez et al., 2008), social network services with geo-tagging (Girardin et al., 2008; Rattenbury et al., 2007), and RFID-enable subway systems (McNamara et al., 2008) . Recent literature has analyzed such datasets to broaden the knowledge on human mobility; however most works only investigated a specific dataset, which reflects only one side of human mobility. Some other works considering several means of transportation have studied travel mode choice in different situations (Mitra, 2013; Hamre and Buehler, 2014; Paulssen et al., 2014).

This paper aims to find daily spatial-temporal ridership patterns, depending on different means of transportation at different locations in Seoul, and to analyze the connection between ridership patterns and location characteristics. This work was triggered by the assumption that such transportation methods exhibit different ridership patterns even in the same locations; thus, a taxi and a subway, quite different from each other, were selected to detect different ridership patterns in Seoul. Determining different connections between the area characteristics and the ridership patterns in Seoul is achieved using the data generated by two different transportation methods in Seoul.

The subway, as a mass-transit service, is the main means of transportation in Seoul. Various research has been performed utilizing the data collected by the electronic payment system equipped in subways to determine area characteristics near subway stations and spatial-temporal ridership distributions (Kim and Lee, 2010; Lim et al., 2012; Shin et al., 2013). Several different purposes and needs of movement, however, can be solved by taxis in metropolitan areas such as Seoul where various and complex economic activities are held. It is because a taxi that can provide services to any location at any time is an undiminished transportation in large cities. For this reason, taxis are

E-mail address: kyoungok.kim@seoultech.ac.kr.

https://doi.org/10.1016/j.jtrangeo.2017.12.003
Received 21 December 2016; Received in revised form 15 March 2017; Accepted 12 December 2017
0966-6923/

alternative to gathering individual trajectories at the city level and many studies have analyzed taxi trip data to establish urban planning and infer human activities, as well as determine the correlations between land use and human mobility in large cities, such as New York, Shanghai, and Beijing, which allowed previous research to obtain rich information on land use and human mobility (Jiang et al., 2009; Liu et al., 2010; Liu et al., 2012a; Liu et al., 2012b; Castro et al., 2013; Shen et al., 2013; Qian and Ukkusuri, 2015; Tang et al., 2015). However, searching for studies that utilize the data generated by taxis in Seoul is difficult because the Seoul Metropolitan Government has begun to open the pre-processed taxi trip data since 2015 on a monthly basis. In addition, previous research usually analyzed a dataset for one of subway and taxi (Kim et al., 2009; Nishiuchi et al., 2012; Lee et al., 2013; Jiang et al., 2009; Tang et al., 2015; Wang et al., 2016) and no study compared the ridership patterns of subways and taxis, even for other cities.

In this study, clustering and classification were chosen as main analysis tools and subway stations were selected as the main spatial areas to be analyzed using the two different datasets. Clustering analysis was applied for detecting groups of regions with similar ridership patterns, which has been widely used for the similar purposes (Liu et al., 2012b; Lee et al., 2013; Zhang et al., 2014; Xu et al., 2016). Through clustering analysis, the stations showing the similar ridership patterns were grouped into the same cluster for the subway and the taxi data, respectively. Then, the clusters from two datasets were compared to distinguish the similar and dissimilar properties. Classification analysis was used to find significant factors affecting the ridership patterns. Here, output target was created based on the clustering results and LASSO (Least Absolute Shrinkage and Selection Operator) logistic regression (Lee et al., 2006) was applied to reveal the effects of the features on the ridership patterns and select important features to predict the types of ridership patterns at once. In this paper, the ridership patterns by taxi and subway were compared around subway stations. The factors that caused different ridership patterns by taxi and subway were determined among several features: land use, statistics related to population, households, houses, companies (the spatial distribution of industries and employment statistics) and others.

The rest of the paper is organized as follows: In Section 2, data sets used in this research are described and then preliminary analysis on them is performed in Section 3. The results of the main analysis, clustering analysis considering ridership patterns related with taxi and subway are presented in Section 4. Finally, implication and conclusion of the paper are given in Section 6.

2. Study area and data preparation

2.1. Preliminary analysis on the public transportation and taxi use in Seoul

Seoul is the capital and most populated city in South Korea. In 2015, 9.9 million people out of 51.7 million people was living in Seoul, which is almost 20%[1]. Seoul is the center of economy, politics, and culture in South Korea, and many businesses, commercial, cultural, and public facilities are concentrated in Seoul. In Seoul, subways play an important role in transportation. In 2014, the modal share rate of subway in Seoul was 39.0%, which is the largest portion compared with other modes, such as bus, car, and taxi. In addition, when the public transportation data[2] collected by the electronic payment system equipped in the buses and subways of Seoul was analyzed, the origins and the destinations of 7959 origin-destination (OD) pairs were both subway stations among the top 10,000 OD pairs of total 3,919,648 unique pairs of OD based on trip frequency[3].

A taxi, as the opponent transit of subways, was considered in this paper due to its distinct properties. This paper focused on areas near subway stations because subways can only be accessed in subway stations. The taxi data collected in 2015 was used for the preliminary analysis to examine taxi ridership in the areas near subway stations. This data contains aggregated trip records of taxis operating in Seoul, which reported the daily time and number of pick-ups (PU) and drop-offs (DO) at specific road points in Seoul. The recording of positions at roads is referred to as ‘link’, each of which represents a 150 m section of the road. The interval of records in time was 30 min; thus, the PU and DO of each link were summarized 48 times a day. A total of 36,455 links were found in the taxi data. When the top 10% of links with high PU and DO were selected, 323 (for PU) and 316 (for DO) links were within 500 m of a subway station and 359 (for PU) and 361 (for DO) links were within 1 km of a subway station. Moreover, the top 10% of links take approximately 53% of the total PU and 46% of the DO. Therefore, subway stations were appropriate research areas.

2.2. Data preparation

2.2.1. Subway and taxi datasets

The subway dataset contains the number of passengers who get on or off the subway stations at the Seoul Capital Area (SCA) in the one-hour interval in 2015 on a monthly basis. A total of 278 stations are located in Seoul among all SCA stations[4]. Two stations, ‘Yangjae Citizen's Forest’ and ‘Cheonggyesan’ stations, out of the 278 were omitted because they are stations for only Sinbundang Line whose records were not open to the public. The subway data separately provides the number of passengers for the same station with different lines. The monthly basis records were converted into annual because the taxi data only provides aggregated statistics over one year.

Furthermore, the taxi data provided daily PU and DO at each link in a 30-minute interval from midnight with every link in Seoul. Two adjacent rows of the taxi data (e.g., 0:00 and 0:30) were merged into one record because the subway data provided the counts in the one-hour interval. In this data, latitude and longitude sequences of each link were provided. The centers of the link sequences were computed and used to determine whether a link was located in a range of 670 m from a subway station[5]. Among 36,455 links, 13,450 (36.6%) remained after filtering. After filtering out links based on location, the PU and DO of links within the 670 m range from the same station were summarized. Finally, the taxi data was converted to the same format as the preprocessed subway data, which provided PU and DO in a one-hour interval over one year for subway stations.

Table 1 summarizes the summary statistics of PU and DO in two preprocessed datasets. In the table, Std. refers to a standard deviation. In the subway data, the mean values for PU and DO were less than the median, whereas the opposite trend was observed in the taxi data. This finding suggests that the subway data was right-skewed, whereas the taxi data was left-skewed, which implies that few subway stations have high usage frequency and few links has low usage frequency.

In addition, Fig. 1 shows the temporal distributions of PU and DO of subways and taxis. This figure shows the total sum of PU and DO for one year. The maximum value of y-axis for the subway data is 10 times greater than that of the taxi data, which reflects that subways takes up a large portion of traffic in Seoul. Two figures from the subway and taxi data reveal quite different patterns over time. In the subway data, two distinct peaks in the morning and evening were observed, whereas no narrow peak was observed for the taxi data. From these observations, it can be inferred that the main reason for boarding the subway is

[1] This information was provided by Korean Statistical Information Service.

[2] This data was provided by Seoul Big Data Campus (https://bigdata.seoul.go.kr).

[3] All transactions collected from 1 April 2016 to 7 April 2016 were selected for the analysis.

[4] The number of stations in this data is different from that of the public transportation data used in Section 2.1 because the public transportation data contains trip records from less subway lines than the subway data.

[5] 670 m is the distance that people generally walk for 10 min (Wansoo Lim, 2006).