Commuter classification and behavior clustering: Beijing use case

Selene Baez Santamaria
s.baezsantamaria@student.vu.nl

Abstract

Public transportation, centered on subway and bus networks, is an data-rich domain that can benefit from data mining and machine learning techniques. The classification of commuters versus non-commuters/occasional travelers can help government, transport management and operators to better target their policies in order to improve the transportation network in large cities. Furthermore, characterizing commuters by behavior clustering can bring deeper insight into their needs and routines as a whole. This project proposes the usage of ensemble models for classification and clustering of public transport users. For this purpose, transit card data will be used, available from the city of Beijing, China.

Contents

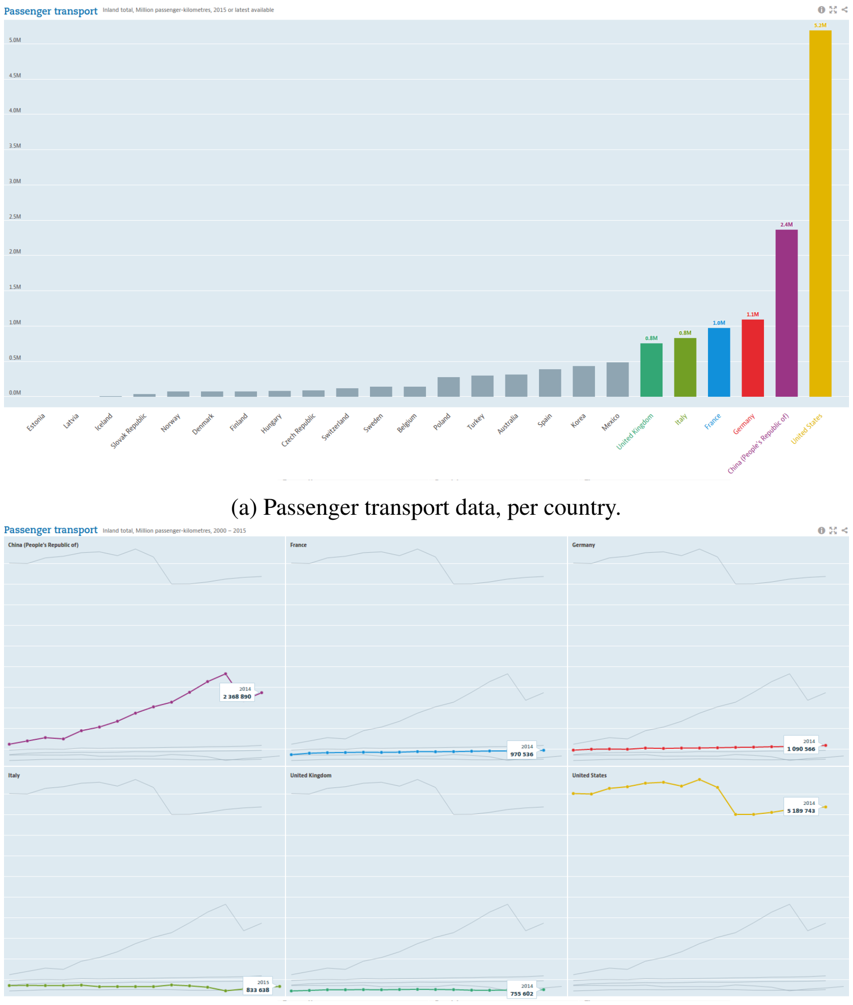
1 Introduction

1.1 Transportation domain

Urban public transportation includes systems that are available for use by anyone in urban areas. Its facilities are commonly composed by buses, subway/metro lines, light rails, tramways, trains and others. As a network, they provide service for the majority of citizens in urban areas. [9]

Figure 1 shows the passsenger transport usage, as million passengers per kilometer. This represents the transport of a passenger for one kilometer. From the top image note that United Stated, China, Germany, France, Italy, and United Kingdom contitute the six countries with the most passenger transport, according to their reported data from 2015 or later [7].

Furthermore, historical data in the bottom image reveals the 15 years behavior for each of the aforementioned countries. Most of the countries show stability, with increase or decrease of less than .10 million passengers for European countries, and .5 million passengers for United States. China, however, shows a trend with steep increase for most of the selected years. In fact, comparing its less than 1.2 million passengers in 2000, China doubled its public transport usage to 2.4 million passengers in 2015.



(a) Passenger transport data, per country.

(b) Historical data for the top six countries with most passenger transport usage.

Figure 1: OECD countries and their passenger transportation data.

The usage of the public transport network as a whole has a significant environmental impact, affecting noise and air pollution. [1].

Public transportation also relates to energetic demand, since its facilities are mostly petroleum or electrical based. In terms of global energy consumption, passenger transportation accounts for about 25% of the total world energy consumption. Furthermore, the tarnsportation sector consupmtion increases at an annual average rate of 1.4% [2]. This may bring further economical implications.

In the last years, smart cards systems appeared, making it possible to track travelers and facilitating the fare collection. Examples are Octopus card in Hong Kong, Oyster card in London, OV-chipcard in The Netherlands and IC card in Beijing, to name a few.

1.1.1 Who are the commuters?

A major proportion of public transport users is represented by commuters. These are regular users of public transit, with regularity in the boarding times/stops.

Driven by a routine, commuters travel back and forth from specific places, commonly represented by home to work/school trips. As commuters use public transit almost daily, or with a regular temporal pattern, the conditions of the public network directly influence commuter's personal life and generally impacts their quality of life [2]. If the commuting experience is bad, daily travel can bring sorrow to users. Bad experiences may include excessively long commuting time, crowded spaces, inconvenient transfers, elevated prices, low-quality of facilities, and others.

Identifying commuters can help in the long-term planning (sustainable) of public transportation. Policies for improving the overall experience and aid urban areas on a large scale.

Transportation follows swarm behavior. Based on individual travels and routines, on a larger scale travelers exhibit peculiar characteristics. Both levels of understanding are crucial.

1.2 The city of Beijing

Beijing special case for urbanization. Number of people is massive [3]

Pollution in Beijing [4]

Beijing has more than 1000 bus routes and 18 subway lines, and it continues to expand every year. This results in more than 28,000 stops combined. [5].

Additionally, Beijing has implemented several Bike sharing systems, with specific dropping stations. [6].

Number of users per day.

By the usage of smart transit cards, collection of payment and transaction monitoring is possible. Buses, subways, bikes and taxis are monitored. Real time and historical data is available. Over 90% of public transit users are smart card holders. [7]

1.3 Motivation

Interdisciplinary study between Artificial Intelligence and Metropolitan Transportation. Introduce data mining techniques to a data rich domain.

Relevance of project on both areas.

[1] references

[2] reference

[3] reference

[4] reference

[5] reference Beijing transportation center

[6] where to get this info

[7] reference

1.3.1 Societal context

Commuters use the public transport network regularly to go to work, school or other follow other routines. They need reliable means of transportation. [8].

The city of Beijing faces a large imbalance between residential and working areas. [9]. Targeting this group brings the largest benefits to the public.

Government, transport management and operators can gain spatial and temporal insight. This insight can lead to tangible results, policies and counter measures increasing efficiency of network, adjustable travel fares used as incentives to relieve peak hours, urban planning for residential and industrial land use, and others [10]

1.3.2 Scientific context

Usage of machine learning of data mining has been limited. Current broadly used method is surveys to reach travelers on individual level and aggregated measurements for gathering their collective behavior. The analysis is usually done with statistical methodology.

Surveys are costly and based on self-report, which by itself has bias problems. Other problems are small population and non-representative samples.

Aggregated methods miss the interactions between individuals that cause the collective behavior.

Technology has reached the data collection point, but has yet to reach the analysis part. Transit cards are capable of recording spatio-temporal information at an individual level over long periods of time. This generates large amounts of data.

Many prediction algorithms available. Constant refinement, state of the art must be applied to real life and large impact situations. Domain experts must focus on analyzing insights and using them, not on techniques for curating and making sense out of raw data.

1.4 Thesis organization

This Thesis is organized as follows:

First we do a literature review for previous work on mining transit data and for specific state-of-the art methodologies. Consequently, we establish the scope and objectives of this project. We continue to describe the methodology thoroughly, including the data and the approach. Following this description, we identify three distinct stages of the project and report their corresponding experimentation. Then, we discuss the findings and gather conclusions. Finally, future work opportunities are explored.

[8] reference: what do commuters care about

[9] reference

[10] reference

2 Literature review

2.1 Data mining on transit card data

Preprocess data by Wang in BJUT lab. [10]

Data mining to identify transit use cycles in Canadian smart card data [6]

Density Based Scanning Algorithm with Noise to classify travelers according to their travel patterns. [5]

Passenger segmentation by K-means clustering [1]

Machine learning for commuters identification. SVM with 94% accuracy. [8]

11 distinct clusters of users with similar activity and demographic attributes [3]

The latest work on the field corresponds to Ma et Al [4]. The objective of their work is to determine a scoring function for travelers that can correctly identify them as commuters, or non-commuters. In their work, they cluster stops using an improved DBSCAN algorithm. They engineer features for representing the frequency in which travelers follow spatio-temporal patterns. Travelers are then clustered according to these features following the ISODATA algorithm. As an output of the clustering, optimal cutoff levels in the scoring function were determined. As a result, evaluating a traveler does not depend on clustering centroids, but only on calculating the commuting score. This, as expressed by the authors, reduces computing time and treats each traveler independently from the others, which is not true for clustering algorithms.

A common practice, as used by [4] and [3] is to divide the day into -hourly or half-and-hour- time bins.

2.2 Classifying and clustering spatio-temporal data

Ensemble methods

Classifiers in the transportation domain

3 Research objective

Objective is to identify and characterize commuters in the city of Beijing by using IC card data. Find patterns in the spatio-temporal data of public transport travelers.

3.1 Research questions

1. How accurately can commuters and non-commuters be identified using an ensemble model? How does this compare to the previous SVM model?
2. What is the minimal set of information needed from IC card data to reach an acceptable accuracy in classification?
3. To what extent is clustering commuters by its behavior informative to transportation specialists?

3.1.1 Definition of terms

A commuter is a public transit user whose IC card data reveals repeatable patterns in time and space over a working week (5 days, Monday through Friday).

A trip is a sequence of IC card transactions, including transfers, with an origin and destination. A trip is also represented as a record in the data, as it will be further explained in Section 4.1

A transfer is a change in transportation mode. Transportation modes include Bus, Subway, and Bike. Transfers can then be: bus-bus, bus-subway, bus-bike, subway-bus, subway-bike, bike-bus, bike-subway or bike-bike. Changes between subway lines are not recorded. [11]

We make the assumption that IC card IDs and users have a one to one relationship, meaning each user has exactly one card and each card is used by exactly one user. [12]

3.2 Scope and structure

[13] The coverage of this Thesis is divided in three main stages:

Part one : classify commuters versus non-commuters. Ensemble model compared to SVM

Part two : the set of features will be revised to disregard redundant information. A second comparison with Tu[8]'s SVM model will be made.

Part three : commuters will be further clustered according to patterns in their behaviors that will emerge from all variables of the IC card data. The clusters will be analyzed and interpreted to find distinctive characteristics that may be judged as useful by transportation specialists.

[14]

[11] transfer possible?

[12] reasonable assumption?

[13] revise this part

[14] better itemize bullets

4 Methodology

4.1 The data

Every record int he data represents a trip performed by a specific IC card. As such, it contains the following data fields:

- Data date: Year, month and day that the trip was made
- Card code: card identification number
- Path link: Mode of transportation. B for bus, R for subway, Y for bicycle. Transfers between modes are shown by a dash. [15] [16]
- Travel time: Time spent in vehicles, measured in milliseconds
- Travel distance: Distance traveled, measured in meters [17]
- Transfer number: Number of changes in travel mode during the trip. Regarding transfers between same travel mode, bus to bus, and bike to bike transfers are counted. Subway to subway transfers are ignored.
- Transfer average time: Time spent in transfer, divided by number of transfers. , Measured in milliseconds [18]
- Transfer total time: Total time spent in transfer, measured in milliseconds
- Start/End time: Time stamp of when the trip started/ended. Date and time with milliseconds precision
- On/Off small traffic area: Integer from 1 to 1911
- On/Off middle traffic area: Integer from 1 to 389
- On/Off big traffic area: Integer from 1 to 60
- On/Off ring road: Integer 1 to 6
- On/Off area: Integer from 1 to 18 corresponding to the district/county for boarding and alighting, correspondingly. See Figure 2 [19] [20]
- ID: record identification number created by joining the following: hour | time stamp of beginning of trip | card code
- Transfer detail: Station name, line number, mode of transportation

The traffic zones (small, middle and big areas) are divided by the Planning and Designing Institute in Beijing[21]. They are specific in different degrees. In general, the division principles correspond to the geopolitical environment and administrative planning, for example roads, villages and others. The 6 ring road and 18 areas are divided by the Beijing Municipal Government [22]. The division is unique in Beijing. According to domain expert Liang Quan, these divisions are sufficiently informative for traffic analysis. [23]

Every day, more than X records are collected. 50, 000 records are sampled every day for a month. [24] In this case, the samples correspond to April, which does not overlap with holidays and has a relatively stable weather thus diminishing the variance between bike and bus/subway traveler preferences. [25]

[15] Example: B-B is Bus to Bus.

[16] calculated column? there are some inconsistencies between this and transfer detail

[17] as measured by route? or start-end stops

[18] calculated column?

[19] reference

[20] include in preprocessing dictionary?

[21] reference

[22] reference

[23] how to quote a person

[24] how much data can we handle

[25] request a proper month data

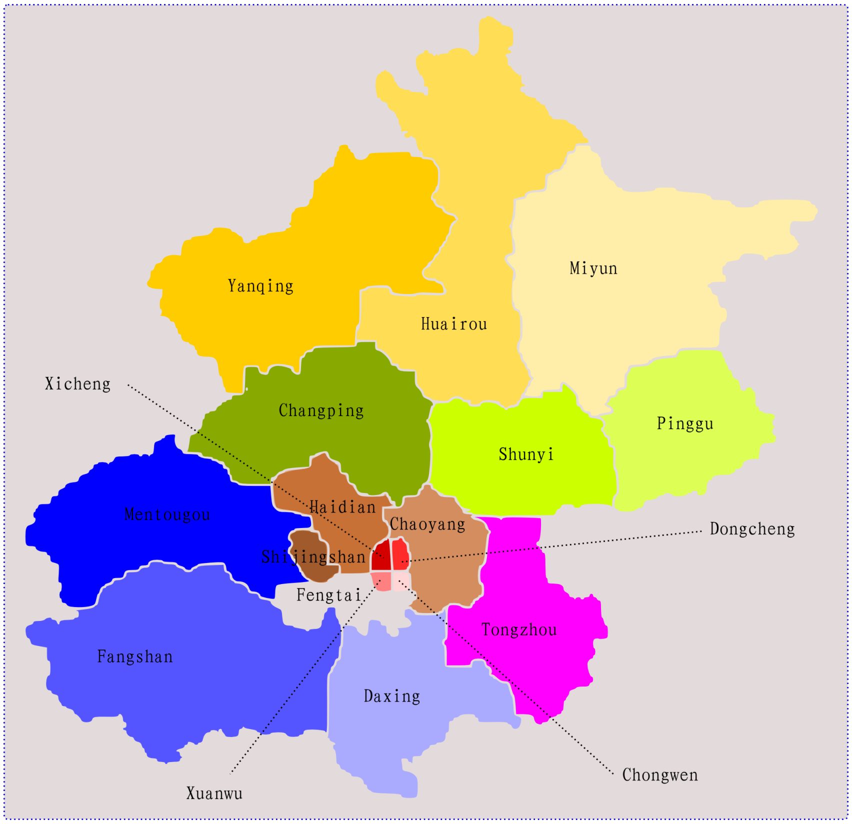


Figure 2: Beijing's Districts and its Counties

4.1.1 Training data

[26] Since we perform supervised learning we need training data for which we know if a record corresponds to a commuter or non-commuter. Such data is expensive and limited since it can only been obtained by asking the users directly if they are commuters or not. Other annotated data is not available, and labeling new records falls beyond the scope of this project. [27]

The current training and validation set consists of data from 2015, collected and validated by Tu [8] [28]. The data is composed by:

- 6439 records of 481 commuters
- 1628 records of 497 non-commuters

For a total of 978 IC card IDs. [29]

4.1.2 Testing data

Testing data is from 2016. More detailed

[26] do we have the correct classes?

[27] if data is not sufficient (although previous work shows it is) I might need to consider annotating some data myself

[28] make sure it was Tu

[29] I got these from Tu, check the parameters are the same as the ones given by Liang or search for IDs in current data

4.2 Data preprocessing

4.2.1 Cleaning

As first step for preprocessing the data, we perform a cleaning where we eliminate records that are faulty, for example:

1. Eliminate records with missing data: 10.9% records eliminated
2. Eliminate records with travel time <= 0: <0.01% records eliminated
3. Eliminate records with travel distance <= 0: 10.5% records eliminated [30]
4. Eliminate records linked to users with insufficient trips: 25.8% records eliminated when requiring at least 2 trips per day.

This leaves 52.6% records available for usage.

Insufficient trips regulated by a user input [31]

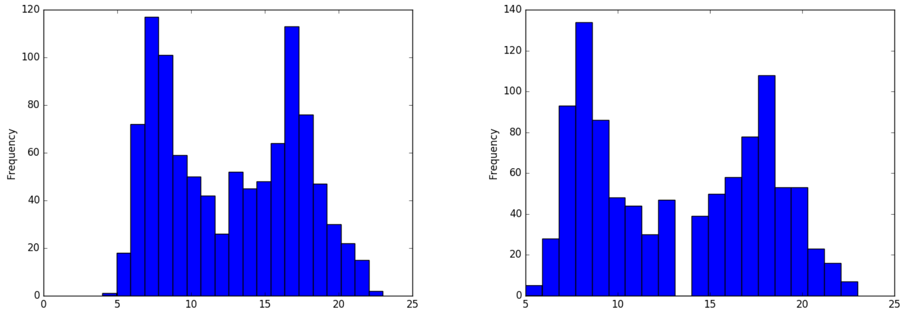
4.2.2 Conversion

Parse route. Translate Chinese keywords.

Dictionary [32]

Example in Chinese -< English -> clean route

Hourly time bins is standard practice in the field [3] [4],



(a) Start hour

(b) End hour

Figure 3: Distribution of start/end hours for trips.1000 records sample.

Figure 3 shows clear morning and evening peak hours.

4.2.3 Standardization

Whitening vs standardization:

Figure 4 shows a clear correlation between travel distance and travel time. [33]

Travel time and distance where standardized by subtracting the mean and forcing a standard deviation of 1 [34] [35]

[30] how do we interpret these?

[31] plot percentage of records and min records needed. Tune parameter

[32] encode chinese in latex

[33] Correlation between time and distance to be preserved?

[34] double check this

[35] replace images once run with all data

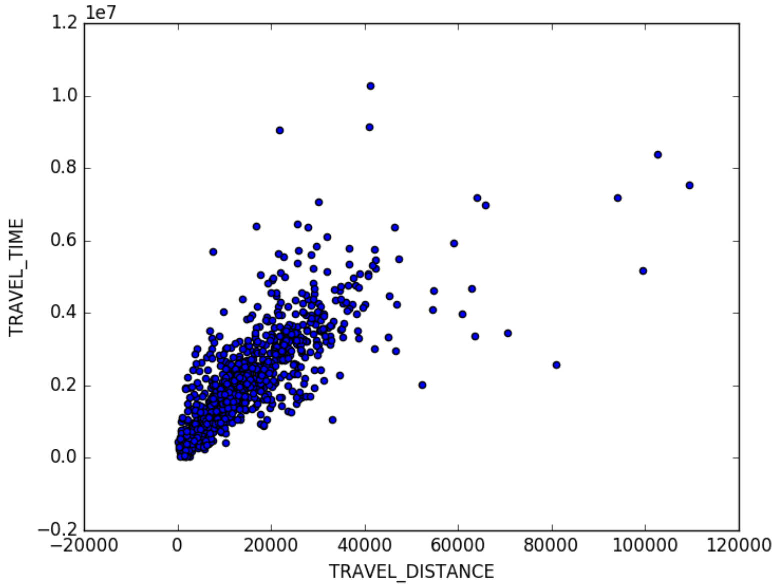
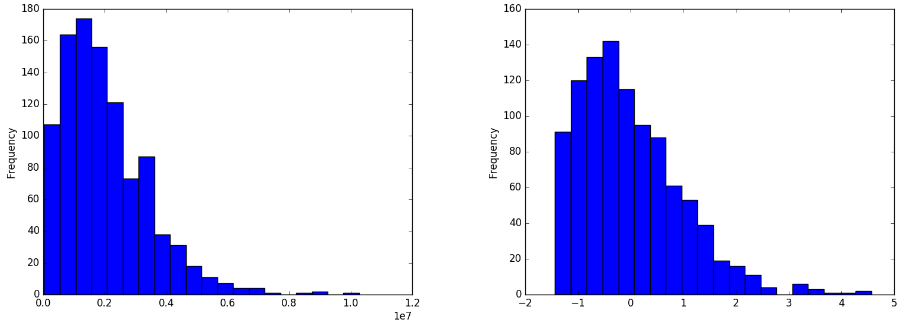


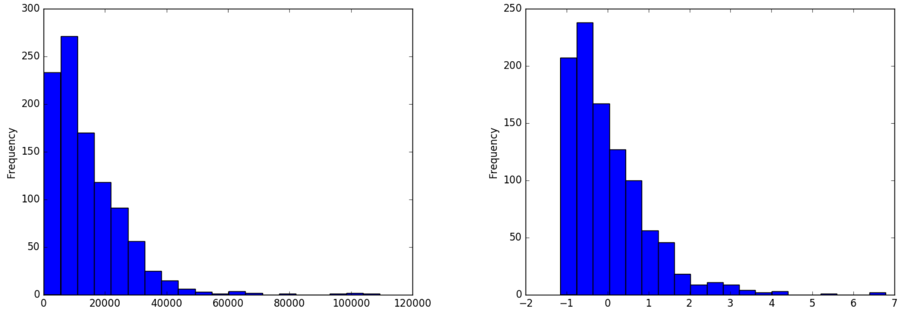
Figure 4: Travel distance vs travel time.1000 record sample.



(a) Original

(b) Standarized

Figure 5: Time distribution before and after preprocessing. 1000 records sample.



(a) Original

(b) Standarized

Figure 6: Distance distribution before and after preprocessing. 1000 records sample.

4.3 Data mining techniques

Coding using Python. Libraries and toolboxes such as pandas, sklearn, theano.

4.3.1 Feature engineering

The temporal factors to be explored are represented by the start/end times, as well travel/transfer time.

The spatial factors to be explored are represented by On/Off areas. [36].

4.3.2 Ensemble models

Ensemble models are chosen because of its robustness and modularity. Starting from two simple classifiers, assembled via bagging, the model can grow larger or more complex as needed and it may be extended beyond the scope of this Thesis Project.

4.3.3 Decision trees and random forests

4.3.4 Neural networks

4.4 Correlation analysis

chi-test

[36] and route lines?

5 Commuters identification

5.1 Hypothesis

As suggested by Tu [8] results, the data is almost linearly separable thus simple classifiers such as decision trees may suffice.

5.2 Model

A first instance of the model will use all available variables in the data as used by Tu [8] for a fair model comparison.

5.3 Experiments

5.4 Results

Accuracy

Confusion matrix

6 Variable evaluation

6.1 Hypothesis

One of the main focuses of the second phase of this thesis is to determine the appropriate level of detail in the area to be taken into account.

Middle area, big area and (small) area overlap. Middle and small divisions have more precision but maybe not needed. On the other hand, big area divisions might not capture the changes for people who live and work/study in the same bis district.

6.2 Qualitative

Exploration: Experts opinion

6.2.1 Interview

[37] We interview Liang Quan as an Transportation domain expert.

- To what extent do people live and work on the same area?
- what level of detail do you think is appropriate?

6.3 Quantitative

Analysis: Correlation

[37] In appendix?

7 Commuters clustering

7.1 Model

7.2 Experiments

7.3 Results

7.4 Expert judgment

8 Conclusion

9 Future work

References

[1] Ashish Bhaskar, Edward Chung, et al. Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3):1537–1548, 2015.

[2] US EIA. Energy information administration (2016), international energy outlook 2016, with projections to 2040. Technical report, DOE/EIA-0484, 2016.

[3] Gabriel Goulet Langlois, Haris N Koutsopoulos, and Jinhua Zhao. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16, 2016.

[4] Xiaolei Ma, Congcong Liu, Huimin Wen, Yunpeng Wang, and Yao-Jan Wu. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58:135–145, 2017.

[5] Xiaolei Ma, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36:1–12, 2013.

[6] Catherine Morency, Martin Trepanier, and Bruno Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[7] OECD. Passenger transport (indicator). `10.1787/463da4d1-en`, 2017. Acessed on 10 April 2017.

[8] Qiang Tu, Jian-cheng Weng, Rong-Liang Yuan, and Peng-fei Lin. Impact analysis of public transport fare adjustment. *Traffic Engineering & Control*, 57(2), 2016.

[9] Vukan R Vuchic. Urban public transportation systems and technology. 1900.

[10] Yueyue Wang. Research on methods of extracting commuting trip characteristic based on public transportation multi-sourced data. *Beijing University of Technology*, 2014.