TU Delft

EXPLORING THE EVOLUTION OF **PASSENGER CHARACTERISTICS** BASED ON **SMART CARD DATA**

A Case Study of Shenzhen, China

QUE JIJIA

Source: https://www.zcool.com.cn/work/ZMzA2MzYzOTI=.html

EXPLORING THE EVOLUTION OF PASSENGER CHARACTERISTICS BASED ON SMART CARD DATA: A CASE STUDY OF SHENZHEN, CHINA

A thesis submitted to the Delft University of Technology in partial fulfillment
of the requirements for the degree of

Master of Science in Transport, Infrastructure and Logistics

by

Jijia Que

24 October 2019

Jijia Que: *Exploring the Evolution of Passenger Characteristics Based on Smart Card Data: A Case Study of Shenzhen, China* (2019)

🅭🅯 An electronic version of this thesis is available at
http://repository.tudelft.nl/.

Graduation committee: Prof. dr. ir. J. W. C. (Hans) van Lint (chair)
Dr. ir. R. (Rob) van Nes (daily supervisor)
Dr. W. (Wei) Pan
Ding Luo

PREFACE

Out of interest in data processing, I chose this study on smart card data. It was a real challenge for me who without any experience. The completion of this master project cannot be separated from the help and support of many people.

First of all, I would like to express gratitude to my committee members. Thanks to Professor Hans van Lint, Rob van Nes and Wei Pan for their professional advice and time spent to help me improve the report. I also want to thank Ding Luo for providing the raw data and technical support.

Secondly, I want to thank my friends. Thanks my friends for patiently listening to my complaints and encouraging me to face up to all difficulties. Especially my boyfriend, he always holds my hands to help me build a solid foundation out of stress. I thank them for making my life in the Nederlands full of happiness.

Finally, I wish to thank my parents, without their financial support and spiritual encouragement, I can not complete the master's degree.

Jijia Que

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 RESEARCH MOTIVATION

Passenger travel characteristics (PTCs) are characteristics that are sought from the information contained in trips, such as travel frequency, travel mode, departure location, and etc. Understanding PTCs can help develop passenger-oriented planning and service policies, so analysis of PTCs is a topic of constant interest to researchers and transportation service providers. However, PTCs vary with the change of many factors, such as network, price policy, service. The process or trend of PTCs change, which is called the evolution of PTCs, are useful for assessing the impact of external factors changes on passengers.

For a long time, the data used in related researches largely comes from traditional collection methods, such as travel surveys. But the traditional collection method is time-consuming and expensive, because the research of PTCs evolution requires travel information during a relatively long period from reliable observers (Chu, 2015). In addition, the data from traditional method may come from biased responses and have a low accuracy due to dependence on interviewees' memory.

With the development of technology, automatic toll collection devices are commonly used in transportation systems, such as automatic fare system (AFS). With AFS, the passenger should pay through the smart card (SC) each time they check-in and check-out In the process, a large quality of individual travel information can be recorded including card ID, boarding time, transaction type, boarding stations, etc., all the information generated after using SC is called SC data. SC data has a great potential to analyze passenger. Because each card has a unique ID, tracking ID can obtain the travel information generated by the card during a certain period of time, so that the travel characteristics of the passenger corresponding to the card can be known. The emergence of SC data avoids the data errors inolved in traditional collection methods and it is relatively inexpensive. And in the case of high AFS coverage, for example, the subway station almost completely uses AFS for payment, and all passengers' travel information will be stored in SC data, which brings a richer and more comprehensive research sample.

Recognizing the potential of SC data, some studies have demonstrated the feasibility of using SC data to explore PTCs evolution, for example Briand, Côme, Trépanier, and Oukhellou (2017); Huang, Levinson, Wang, Zhou, and Wang (2018); Viallard, Trépanier, and Morency (2019); Wang, Chen, Wang, and Huang (2018). In these studies, the exploration of the evolution of PTCs is based on the transaction data within the analysis timeframe generated by a fixed passenger group, that is, a sample data is determined. There are two main methods to explore evolution. The first method first determines representative PTCs based on all transaction data in the sample. The analysis timeframe is then divided into consecutive periods that do not overlap, and the transaction data of the passengers in each period is selected. Based on the data for each period, explore each passenger's characteristics at different periods meets which representative PTCs. Finally, by tracking the representative PTCs of each passenger during each period, the PTCs evolution of the sample is known. The study using this method is Briand et al. (2017), and the benefit is that the difference in evolution from different passengers can be known. Its limitation is that the representative PTCs calculated here are based on all sample data, that is, they are unchanged during the analysis timeframe, and the evolution of PTCs exhibited by all sample passengers cannot be observed.

The second method is to first divide the analysis timeframe into consecutive periods that do not overlap, and find the PTCs in each period. Learn about evolution by comparing PTCs from different periods. The PTCs for each period obtained by this method are all characterized by the appearance of all passengers involved in each period. Studies in this method are Chu (2015); Huang et al. (2018); Viallard et al. (2019); Wang et al. (2018). Such an method allows researchers to understand the different PTCs of a fixed group of passengers at different periods, but cannot explore the differences in evolution between different passengers within the group.

In general, it is feasible to explore the evolution of PTCs through SC data, but no relevant research considers the changes in PTCs exhibited by all sample passengers, and also considers the difference in PTCs evolution between different passengers. Understanding the evolution of PTCs across the entire passenger group and the evolution of different passenger PTCs in the group can help to know the impact of changes in policies or services on passengers from different levels.

1.2 RESEARCH OBJECTIVE

This thesis aims at exploring the evolution of PTCs and the difference of the evolution among passengers. The analysis is accomplished based on the SC card data from Shenzhen Metro, which provides complete historical records of all passengers' transactions in December 2012, December 2013 and December 2014. The objective of this research is achieved by answering following research question:

What is the evolution of passenger characteristics based on SC data?

The main question is solved by dealing the three sub-questions:

- Based on SC data, how to describe the characteristics of passengers?
- How to identify the change of PTCs?
- What are the changes in PTCs during the analysis timeframe?

1.3 RESEARCH OUTLINES

The main structure of this study is as follows:

- Chapter 2 is literature review. According to previous research, the indicators used to describe passenger characteristics are summarized, and the scientific gap is also discussed here.

- Chapter 3 is methodology, which introduces how to describe PTCs from SC data and how to find the change of PTCs
- Chapter 4 is the research background and data description, which mainly includes the research area, the basic situation of the Shenzhen Metro during the period 2012-2014 and data type.
- Chapter 5 shows the results of PTCs for passengers in different periods and discuss the PTCs change.
- Chapter 6 summarizes the main findings. In addition, the limitations of this study and future research directions in this field are discussed.

2 LITERATURE REVIEW

There is a large number of researches that analyzes passengers through smart card data. Pelletier, Trépanier, and Morency (2011) summarizes some related studies and states that SC data is helpful on passenger characterization and classification without having personal information. By passenger characterization, it is possible to obtain characteristics about passenger travel behavior, that is, obtain PTCs. The passenger classification is to distinguish the passengers, which is helpful to understand the difference between the passengers. Therefore, this chapter would conduct literature review to understand how passenger characterization and classification were carried out in previous studies. In addition, the literature focusing on PTCs evolution would be summarized to understand the method they have taken in exploring evolution. These would bring inspiration to subsequent research.

2.1 PASSENGER CHARACTERIZATION BASED ON SC DATA

From the existing research on analyzing passengers through SC data, the researcher's description of PTCs is from different perspectives according to the research requirements. Some of the perspectives commonly used in the literature to describe PTCs are introduced below, and from these perspectives, which indicators can be used for quantification. Regarding the choice of indicators, they are different because the research purposes and data involved in the literature are different. Here are some common indicators.

- **Temporal perspective**
  With different lifestyles, PTCs varies with time. Transaction time is an important information contained in the SC data, so investigating this dynamic from temporal perspective become a possible choice.

  In the study of Ma, Liu, Wen, Wang, and Wu (2017), he explore the temporal characteristics for commuter by number of traveling day and number of similar departure times for a week. From the two indicators, he found there is a difference between commuter and non-commuter, and he concluded that the two indicators can be used to distinguish the temporal characteristics of commuters and non-commuters.

  In a study based on SC data from Beijing metro, travel frequency is used to classify commuter and non-commuter (Huang et al., 2018), which is number of trips performed per week. The author proves that passengers with different travel frequencies, their characteristics can be clearly distinguished. So, trip frequency can be an indicator to roughly separate passenger with different characteristics. As a data derived by aggregating the information of several trips, trip frequency would ignore the details contained in these trips, such as the difference in temporal and spatial perspective, and the order of these trips over time.

Both studies above are based on one week. From temporal perspective, it also can be discussed in a shorter time scale, like within one day. The time is divided according to the prescribed interval, for example, 24 hours a day, divided into 1-hour interval (Bhaskar, Chung, et al., 2014), and then the temporal characteristics identified by the time selection. The smaller the time interval divided, the more detailed the travel pattern is described, and the greater the amount of computation required. But the limitation of this method is that some travel behaviors near the time division point should belong to the same temporal characteristics, but in this method, it may be treated as different. For example, when people start trip at 7:59, they are considered to be in different temporal characteristics compared with starting at 8:01, because the two times are divided into different time periods by 1-hour intervals, but they are actually the same in temporal perspective.

In addition to the above indicators, time-related indicators include ridership (Agard, Morency, & Trépanier, 2006; Chu, 2015).

- **Spatial perspective**
  Since the transactions in SC data include trading locations, it is possible to explore passengers' spatial characteristics. One example is the regularity of station, which is used to find the home and job station according to the most used station (Huang et al., 2018). After knowing the home/job station of each passenger, the passenger distribution on home/job station can be visualized combined with the geographic information of each station. However, the drawback of this method has been recognized (Lee & Hickman, 2013). When passengers can choose multiple stations to take a metro to a certain place in the vicinity, no matter how they choose, this should belong to the same. But in this method, because it is divided according to the boarding station, the result will be treated as different.

  In addition to using trading location directly, some indicators calculated in combination with other databases can also be used to describe spatial characteristics, such as selected route and trip distance. SC data only includes origin and destination station, so the route selected needs to be inferred in combination with the network information (Ma et al., 2017). Trip distance also cannot read from SC data, which can be inferred based on travel time or calculated based on the selected route. By using trip distance, the range of passenger's movements can be known (Hasan, Schneider, Ukkusuri, & González, 2013).

- **Activity perspective**
  In some studies, activity is another perspective used to describe PTCs.

  Activity duration is an example. It refers to the length of time a passenger has activities at a destination. Different lengths of time reflect different types of activities. For example, the time at home after work tends to take longer than the time when working outside. Ortega-Tong (2013a) calculated the longest activity duration of each passenger based on the data for one week, and found that most passengers' longest activity duration on the working day is 8-9 hours, while on non-working days it is 1.5-3 hours.

Combined with activity duration and some other information, such as home/job station, researchers infer different activity status of passengers. With knowing the type and duration of activities, it is possible to capture the activity sequence within which each trips occurs, some studies chose to describe and distinguish PTCs based on activity sequence (Goulet-Langlois, Koutsopoulos, & Zhao, 2016). The advantage of this method is that it can help to understand the connection between each trip. For example, after Friday's work, it may not go directly to the home, but to carry out some leisure activities. However, this method is based on SC data to infer the passenger's activity. The accuracy of the inferred result should be verified by other data, such as questionnaires.

- **Other perspective**
  In addition to the three perspectives mentioned above, there many related researches attempt to interpret PTCs from richer perspectives.

  In SC data, card type reveals passenger's social attributes, such as age and occupation, so this is also possible to find sociodemographic characteristics (Agard et al., 2006; Briand et al., 2017). But this is mainly used to distinguish between special groups, like students or the elderly. For most general groups, the type of SC fails to capture their differences. In addition, different types of cards may reflect different fare policies, which is also a key factor in determining passenger travel behavior.

  SC can usually be used in different modes of transportation, such as subways, buses, trains, so the passenger's travel mode selection can also be known through SC data.

Table 2.1 lists different discription perspectives and related indicators mentioned above.

Table 2.1: Different perspectives and related indicators for passenger characterization

Temporal perspective	number of traveling day, departure time, trip frequency, travel time, ridership
Spatial perspective	boarding station, selected route, trip distance
Activity perspective	activity duration, activity type
Other	card type, travel mode

Actually, there are certain deficiencies in describing from each perspective, so in order to make full use of the potential of SC data and a more complete understanding of PTCs, research is often described from multiple perspectives and using several indicators. Table 2.2 summarizes the literature describing passengers from which the perspective mentioned in this chapter.

Table 2.2: Review of studies on passenger characterization based on SC data for different description perspective

Author/Year	*Temporal perspective*	*Spatial perspective*	*Activity perspective*	*Other*
Ma et al. (2017)	departure time, number of traveling days	boarding station, selected route		
Bhaskar et al. (2014)	departure time	boarding station		travel mode
Hasan et al. (2013)		boarding station, trip distance	activity duration	
Goulet-Langlois et al. (2016)			activity duration, activity type	
Huang et al. (2018)	travel time	boarding station		
Agard et al. (2006)	trip number			card type
Briand et al. (2017)	trip frequency	boarding station		card type
Viallard et al. (2019)	ridership			
Chu (2015)	trip number	boarding station		
Wang et al. (2018)	ridership	trip distance		

2.2 PASSENGER CLASSIFICATION

The passenger classification is helpful to have a better understand of passengers (Bhaskar et al., 2014). There is a massive population contained in SC data, it would be complexed to analyze every passenger within the dataset. But if the behavioral characteristics of the entire passenger group are observed only from the level of aggregation, this completely ignores the difference between passengers. For researchers, passenger classification help researcher only needs to analyze the characteristics of each group of passengers based on the grouping results. While reducing the complexity of the analysis, it is also possible to understand the difference between different groups of passengers. For operators, after understanding the needs of different categories of passengers, they can develop corresponding services to cater passenger. This study is mainly to better reflect the difference in the evolution of PTCs between passengers through the passenger classification. So next, how to classify passenger based on SC data in previous studies would be introduced.

Passengers classification can be based on personal characteristics, which can be indirectly found from card type (Briand et al., 2017). Briand et al. (2017) divided the passengers into 11 categories based on whether thier SC is bound to the bank card, whether the SC can be used for the express route, whether it is a student card, etc., and successfully observed the difference in characteristics exhibited by each type of passenger from a temporal and spatial perspective. However, in SC data, in order to reduce the data size, different card types are usually represented by different numbers. To classify passengers by card type, it is first need to understand the meaning of these numbers, which means that data other than SC data is reqiured.

Another way is to classify according to different travel characteristics. For example, Goulet-Langlois et al. (2016) divides passengers with similar activity sequences into groups. Huang et al. (2018) distinguishes between commuters and non-commuters based on the trip frequency of one week. As can be seen from these two examples, the characteristics used to classify passengers may be the activity sequence inferred from SC data, or the data that can be directly obtained as the trip frequency. In terms of accuracy, the information that can be obtained directly from SC data is more accurate, but the type of these information is limited. For characteristics inferred from SC data, although its accuracy is yet to be verified, it provides a richer perspective on difference between passengers.

2.3 EVOLUTION OF PTCS

The longer it takes for SC data, the more difficult it is to collect and store. The data spans that many researchers can use are relatively short and cannot be used to study the evolution of PTCs (Agard et al., 2006; Bhaskar et al., 2014; Morency, Trepanier, & Agard, 2007). Some studies use data involving a large time span (Deschaintres, Morency, & Trépanier, 2019; Goulet-Langlois et al., 2016; Morency et al., 2007), however, in those studies, their focus is not on the changes in PTCs during the analysis time, but on the characteristics of passengers throughout the period as a whole. Excluding the above situation, there is not much research about the evolution of PTCs. Through Google scholar, the following would introduce all relevant research retrieved from 2015 to the present.

Briand et al. (2017) has SC data for five consecutive years. The author chooses a representative day from each year, for a total of five days. The five-day SC data constitutes a research sample. Then ten representative time activities characteristics were obtained from the whole sample data. Based on the data in first year in the sample, the PTCs for each passenger in the first year are obtained. According to the most similar representative characteristics with their first year PTCs, all passengers are divided into ten groups. By repeating the above steps, the grouping situation of passengers every year can be known. Then, author explores PTCs evolution by tracking the group each passenger belong to in each year. In this study, the change in PTCs per passenger can be understood, so the difference in evolution between passengers can be reflected. But the ten representative characteristics obtained at the beginning are unchanged, which is summed up from the five-day SC data of each passenger. But in fact, if only looking at the data for the first year, different representative characteristics may be obtained. In addition, only one-day data in a year is selected to analysis. In this relatively short period of time, the change in the PTCs may come from some accidental events.

Briand et al. (2017) analyzes the temporal characteristics of passengers, while Huang et al. (2018) studies the spatial characteristics of passengers. In the study of Huang et al. (2018), the authors identified the commuter by a relatively high weekly trip frequency and their home station (station near the residence) and work station (station near the workplace) from SC data. According to the changes in the home stations and work stations of these commuters within seven years, they were divided into four categories. Afterwards, the authors analyzed the changes in average travel time and housing expenditure for each type of passenger during these seven years. That is to say, this research focuses not only on the characteristics of passengers' travel, but also on the connection between passengers' spatial characteristics and other urban factors, which requires the support of other relevant data.

Both of the above studies are aimed at the evolution of the year-to-year, and next are some of the studies involving day-to-day and week-to-week evolution.

The study of Viallard et al. (2019) explored the week-to-week evolution of the characteristics. The PTCs here are described by the number of trips per passenger per day during a week. The purpose of this description is to understand the difference in passenger travel needs per day. The authors classify passengers according to the data in first week and obtain representative initial characteristics, which is discribed by number of trips per week. Then, based on whether there is a policy holiday and school break next week, adjust initial representative characteristics to the new representative characteristics. In this study, the authors study evolution through changes in representative characteristics each week, but the limitation is that this is mainly applied based on the data in last week. If the data of a week is missing, accuracy cannot be guaranteed. In addition, the new week's representative characteristics are adjusted according to that of the previous week, so once the number of representative characteristics is determined, it will remain the same.

The study of Chu (2015) conducted in four levels to find the change in passenger's travel characteristics, including within-day, day-to-day, seasonal, and year-to-year. This study understands the evolution by comparing the passenger's mobility and activity location. The author selected some cards for research, and the transactions of these cards formed sample data. The PTCs are discribed based on the data at each year/week/day generated by all cards in sample, that is, it could not understand the difference between passengers. But the authors discussed how to select sample data suitable for longitudinal analysis from the raw data, which provided inspiration for the subsequent data preprocessing in Section 3.2.

Different from all the above studies, Wang et al. (2018) explore the differences in PTCs before and after the known external factors change (price policy adjustments). From the perspective of temporal and spatial, the author discusses the difference between the PTCs by using indicators - travel demand and travel distance. In this study, because it is possible to accurately know when and what external factors change, the time and possible changes in the evolution of PTCs became easier to infer. But for most cases, it may not be able to know the relevant information about external factors, especially when studying some old data.

After literature review, it is found that the exploration of PTCs evolution in existing research can be divided into two types. One is to group passengers according to the characteristics that are found from all sample data, and then study the belonging group of passengers in different periods, evolution here is represented by member switching (Figure 2.1); the other is to compare the overall characteristics of the target passenger group in different periods, that is, aggregate characteristics changes (Figure 2.2).

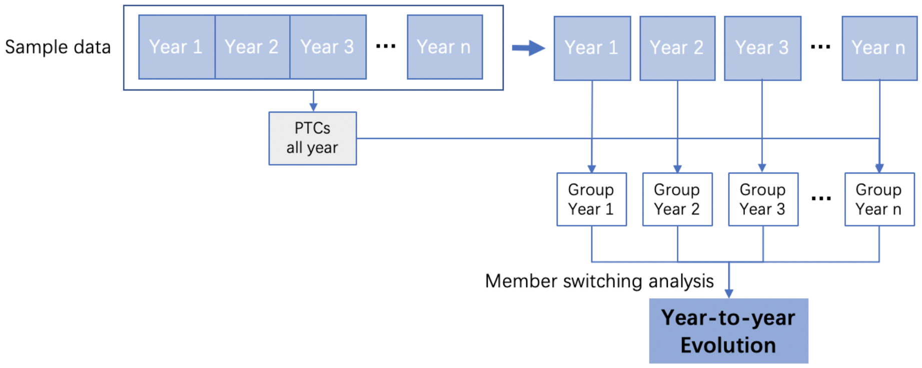


Figure 2.1: Evolution exploration method: member switching analysis (year-to-year for example)

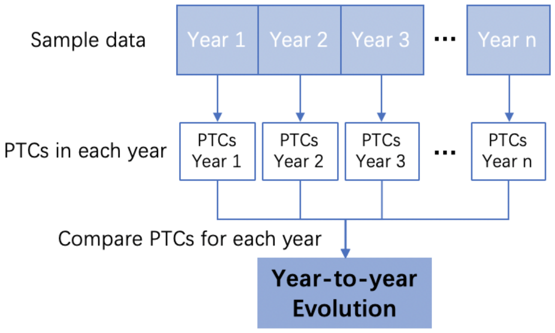


Figure 2.2: Evolution exploration method: aggregate characteristics change (year-to-year for example)

Table 2.3 summaries the research objects, the period of comparison, and the types of evolution of exploration of the studies mentioned before. The indicators used to discribe the PTCs in these studies are shown in Table 2.2.

Table 2.3: Review of studies on PTCs evolution

Author/Year	*Research object in research period*	*Comparison time period*	*Evolution type*
Briand et al. (2017)	All cards	Year-to year	Member switching analysis
Huang et al. (2018)	Commuters	Year-to year	Aggregate characteristics
Viallard et al. (2019)	All cards	Week-to-week	Aggregate characteristics
Chu (2015)	Continuous use cards	Within-day, Day-to-day, Seasonal, Year-to-year	Aggregate characteristics
Wang et al. (2018)	All cards	Before and after the extral factor change	Aggregate characteristics

2.4 SCIENTIFIC GAP

Combined with literature studies on PTCs evolution, the following scientific gaps can be found:

- Some existing studies analyze the changes in the overall characteristics of all passengers and do not see the differenece between passengers' evolution (Chu, 2015; Wang et al., 2018). The study of Briand et al. (2017) can explore the difference in passenger characteristic evolution by focusing on member switching between groups. But in this research, the characteristics of each group are based on all the data from seven years, the dynamic changes of all passegners in seven years have not been explored, and the study by Viallard et al. (2019) also has the similar issues.
- Some changes in PTCs may be caused by extreme weather or other sudden factors, and then recovered and would not be maintained for a long time. For longer time scale, such as the evolution of year-to-year, it is more important to consider the regular travel characteristics that passengers reflect within a year, because this is the characteristics that best represents this passenger. But in existing studies, the effects of those irregular behaviors to PTCs have not received enough attention.

Therefore, in this study, in addition to exploring the evolution of a group of passengers by comparing the PTCs in different periods, it would also show the difference of the evolution among thoes passengers, while trying to reduce the impact of irregular changes on the results.

3 METHODOLOGY

In this chapter, the methodology used in this thesis would be introduced. First, the entire analysis process will be presented, and then the principles on which the data preprocessing and passenger classification are based will be explained in detail. Passenger characterization will explain why passenger characteristics are described from the temporal and spatial perspective. Finally, explain how use longitudinal analysis to explore evolution.

3.1 ANALYSIS PROCESS

Figure 3.1 shows the entire analysis process of this paper, which is mainly divided into four steps and are represented by squares in the figure. The raw data cannot be directly used in the analysis, so the data is first preprocessed by step one. Then, in the second step, all passengers are grouped in order to better observe the difference between the them. After the two steps, the research samples in different periods can be acquired. Then, from the temporal and spatial perspective, the selected indicators quantify the PTCs. After obtaining quantitative results of PTCs at different times, the evolution of the PTCs will be found by longitudinal analysis. The four steps are described in detail below.

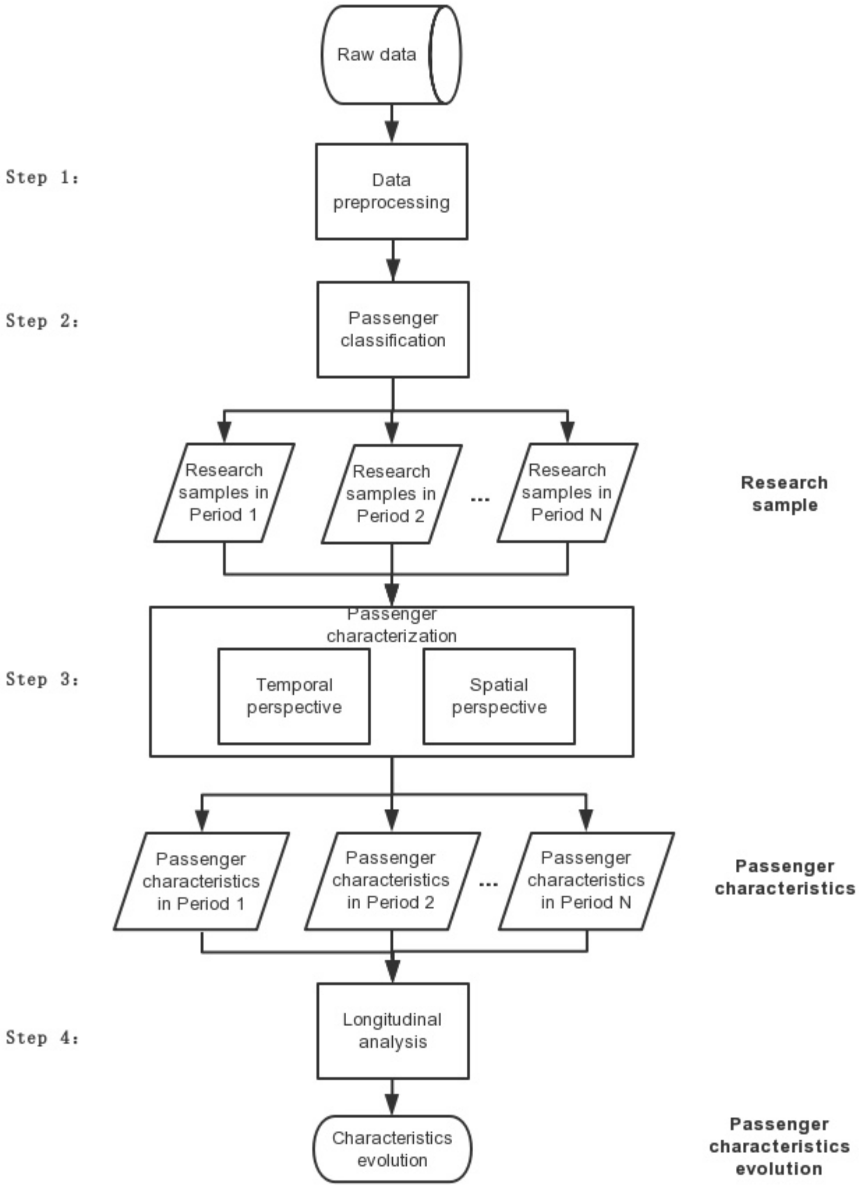

Raw data
Step 1:
Data preprocessing
Step 2:
Passenger classification
Research samples in Period 1
Research samples in Period 2
...
Research samples in Period N
Research sample
Passenger characterization
Step 3:
Temporal perspective
Spatial perspective
Passenger characteristics in Period 1
Passenger characteristics in Period 2
...
Passenger characteristics in Period N
Passenger characteristics
Step 4:
Longitudinal analysis
Characteristics evolution
Passenger characteristics evolution

Figure 3.1: Analysis process

3.2 DATA PREPROCESSING

There may be incomplete data in the raw data, such as the transactions about one trip, may only have check-in record have no check-out record. In addition, there may be errors in raw data. For example, some records may record incorrect ID for station, which makes it impossible to identify at which station the recording occurred. Incomplete data and erroneous data cannot use for analysis, which need to be deleted.

Another part of the problem that needs to be solved in data preprocessing comes from the research objective and research needs of this thesis. This thesis aims at analysising passengers, but the transaction data generated based on card in the SC data. Therefore, for the sake of research, the first assumption is proposed here:

Assumption 1: One card only corresponds to one passenger.

Therefore, in the data preprocessing stage, all relevant transactions will be sorted out by the ID of each card, which is the transactions of the passenger corresponding to this card.

After obtaining passenger-based transaction data, a sample of data suitable for longitudinal analysis will be selected. Reference is made to the research of Chu (2015) about extracting longitudinal observations from the SC data, which is designed to obtain a consistent sample for the analysis timeframe. The specific method is to obtain the active period for each card according to the first transaction time and the last transaction time after determining the analysis timeframe, and only the card that its active period covers the whole analysis timeframe can be retained. The limitation of this method is that the selected card may only have two transaction records with a long-time interval. The active period obtained by the two records covers the analysis timeframe, but actually no other transaction occurs during that timeframe. Especially for discrete analysis timeframes, like a timeframe consisted by January 2012, January 2013, and January 2014. Such methods may obtain some cards have transactions in January 2012 and January 2014, but there is no recording in January 2013, resulting in the inability to analyze PTCs during the period. In order to avoid such a situation, in the case of discrete timeframes, the sampling method of Chu (2015) is improved to only the cards that have transactions data available for all periods in analysis timeframe can be retained.

Also take the analysis timeframe consisting of three months of January 2012, January 2013, and January 2014 as an example. Only the cards with transaction records in these three months can be retained. For the case of continuous timeframe, this method can also be applied by dividing the time involved into several small time periods. Therefore, there are three steps in data preprocessing, which is shown in Figure 3.2.

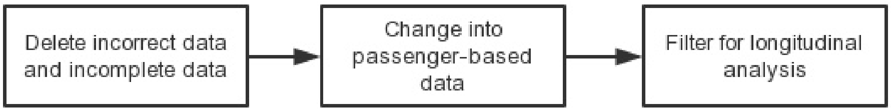

Delete incorrect data and incomplete data
Change into passenger-based data
Filter for longitudinal analysis

Figure 3.2: Steps in data preprocessing

3.3 PASSENGER CLASSIFICATION

Passenger classification better reflect the difference between passengers. In Section 2.2 how to classify passengers have been discussed. In this study, the indicators used for classification will be selected based on the following two points:

- This indicator can be obtained directly from SC data.
- The classification of passengers by this indicator does not require interpretation from other data.

Among them, the first point is to obtain more accurate indicators. For example, activity sequences are inferred from SC data, and errors exist. The second point is to consider the difficulty of obtaining data. Data like card type, in SC data, usually uses different numbers to distinguish different type of cards. But what is the meaning behind the number, is not the information contained in SC data. The more different data sets involved, the more difficult it is to obtain them.

Considering the above two points, the passenger classification indicator used in this paper is trip frequency. The trip frequency of each passenger can be obtained by calculating the number of transactions involved, and the passengers can be classified according to different frequencies. Studies have shown that passengers with different trip frequencies have different characteristics (Huang et al., 2018). But background conditions such as passenger habits, road network and operation conditions are various in different SC data. Therefore, there is no standard for how to divide. It is necessary to first understand the SC data before classifying passengers. In addition, there are many ways to classify, because the focus of this study is not on passenger classification, so a simple method of direct classification based on a single indicator is adopted. It should be noted that passengers need to be reclassified once for different periods involved in the analysis timeframe. After all, even if the travel frequency is the same, the characteristics of the different periods may be different, and the passenger's travel frequency also changes. In this way, with classifying passengers in each period, not only can understand the different characteristics of groups of different travel frequencies in different periods, but also can understand the changes of members between groups.

3.4 PASSENGER CHARACTERIZATION

After the passenger classification, the PTCs of each group in each period of analysis timeframe would be described separately by indicators from temporal and spatial perspective. The reason for choosing these two perspectives is that time- and space-related indicators can often be obtained directly from the SC data, and multi-perspective descriptions provide a more complete picture of PTCs.

3.4.1 Temporal perspective

There two indicators related to temporal perspective are selected here, daily ridership and departure time. They quantify the PTCs in different time scale.

- **Daily ridership**
  The Daily ridership, whose unit is day, calculates the total number of all trips generated per day during the operating hours. The most direct response of daily ridership is travel demand. After obtaining the daily ridership for a certain period in timeframe, it is possible to find the regularity of the passengers associated with it, which is useful in distinguishing between regular passengers and irregular passengers, as well as differences between working days and non-working days (Huang et al., 2018). In addition, changes in external factors can be found through some unusual daily ridership. For example, when extreme weather occurs, daily ridership will show a significant decline (Zhou et al., 2017).

  Here, the choice of daily ridership is mainly to observe the difference between the passengers with different trip frequencies and the demand of passengers on working and non-working day.

- **Departure time**
  Departure time has a shorter time scale than daily ridership, in units of 30 minutes. This indicator calculates the most frequent departure time for each passenger's first trip per day during the observation period and for the last trip. So, the entire operation time (06:00-24:00) can be divided into 36 time segments. Based on the departure time of each passenger's first/last trip each day, the time segment they use most often can be known. This option avoids the impact of unusual departure times.

  Here, the reason for discussing the departure time of each passenger's first trip and the last trip respectively is that, different departure time indicate different travel purpose, like commuters need to have they first trip before 9:00 for arriving working place on time, but for housewives, they have a more flexible departure time for daily shopping. For the last trip, passengers who depart later than 17:00 are more likely to be commuters because this is the time they back home from work. And in order to explore the consistency during observation period, this indicator would calculate on working day and non-working day respectively.

  Therefore, for passengers in each group during each period in analysis timeframe, there are four departure times would be calculated, that is departure time for first trip on working day, last trip on working day, first trip on non-working day and last trip on non-working day. But for passengers who only travel once in a day, this trip is both the first trip and the last trip of the day, and the impact will be discussed in Section 3.6.

  In addition, at the time of sample selection, it is based on whether the card corresponding to the passenger has a transaction record in each period. Here, the departure time needs to calculate according to the daily transactions in each period. It may happen that the passengers did not travel every day during the period. Therefore, if the passenger distribution result in each time interval is expressed as the number of passenger in each group observed during working/non-working days, the distribution difference due to the difference in the number of passengers may occur. What is more needed here is to compare the different distributions of passengers' departure time choices in different groups. The final result will be expressed as a percentage of passenger in each group observed during working/non-working days. This will eliminate the impact of the difference in observed passenger numbers.

Considering the calculation time, the unit is 30 minutes in this thesis, but if higher accuracy is required, the unit can be shortened

3.4.2 Spatial perspective

There two indicators, trip distance and origin/destination, is selected to describe PTCs from spatial perspective.

- **Trip distance**
  Trip distance has been shown to be linked with mobility and how accessible to activity locations (Hasan et al., 2013; Ortega-Tong, 2013b). Here, be calculating the mode of geometric distance for all trips generated by each passenger during the observation period, it is possible to have an insight into the distribution of trip distance among each group.

  The mode value chosen here instead of the average value is to reduce the impact by extreme value caused by the accidental change of travel behavior, which is especially important for discrete data such as travel distance. For example, a passenger has a total of five trips during a certain observation period, four regular trips are 3 km, and an accidental trip with the distance of 20 km, so the average value is 6.4 km per trip. The mode can exclude the accidental trip and the result is 3 km.

  In addition, trip distance cannot be read directly from SC data. The data that can be obtained is only the check-in/check-out of the passenger at which station. In the study of Hasan et al. (2013), the check-in and check-out time for on trip are found first to calculate the travel time, and then multiply by the average speed of the train or bus (transport mode related to the SC data) to obtain the travel distance. This method is simple, and the calculation time is short, but it is not accurate enough. The reason is that the travel time obtained also includes the time for the passenger walking in the station, and the train/bus is not running at a constant speed. Eventually it may result in the different travel distance even between the same start and end station. In order to obtain more accurate results, the geometric distance is calculated here in combination with the network data. In the case of no network data, because of the popularity of geographic data, it can be obtained from public maps, such as OpenStreetMap.

  The specific way is that, the transaction data in sample is first converted to a trip-based format (Table 3.1) based on the two transactions data (one for check-in, one for check-out) related to one trip. From Table 3.1, only the starting and ending stations of the passenger can be known, but the specific route selection of the passenger cannot be determined. In order to calculate the trip distance, the approach taken here is

Table 3.1: Example of trip-based data

Card ID	*Date*	*Time*	*Start_station*	*End_station*
20000513	3	"19:52:04"	"1268002000"	"1268004000"
20000513	5	"19:29:25"	"1268002000"	"1268004000"

to assume that the passenger will always choose the shortest route. Combining the GIS data of the network with the starting and ending stations of each trip in Trip data, the Dijkstra algorithm (Dijkstra, 1959) is used to determine the shortest path, and the length of the shortest path found is the trip distance. After obtaining the distance for all trips, the mode of trip distance for each passenger can be calculated then. Here, the working days and non-working days will be discussed separately to better explore the different travel distances of passengers in these two periods.

Finally, when comparing the travel distance distribution of each group of passengers, the distribution value will also be expressed as a percentage, the reason is consistent with the departure time.

- **Origin/destination area**

  The origin/destination area obtained by the most used frequently departure station for each passenger's first/last trip per day during the observation period. Most used frequently station can indicate the locations where passengers are often having activities, it is an important part for passenger mobility. The reason why chooses most used frequently station is to eliminate the effects of irregular change.

  People's activity areas differ between working days and non-working days, as evidenced by the difference in the number of visitors to each station on weekdays and non-working days in the study of Gong, Lin, and Duan (2017). In order to understand such differences, the working days and non-working days will be discussed separately when calculating the destination area. The reason for discussing destination area on non-working day and working day seperatel is that, the departure station of the last trip of each day is more likely to be different between the working day and the non-working day than the first trip. For example, the last trip of the commuter on working day usually departures from the workplace, while on non-working days, they are more likely to start from leisure activities. But their first trip of one day is usually from home. Another reason is to simplify the research.

  In addition, for the passenger with low trip frequency, the number of trips that each one can count may be only once in a day. Here, in order to ensure accuracy, the starting station of their first trip is selected to decide the origin area, and the end point of the first trip is for the destination area.

  After knowing the station involved in the origin/destination area, the area around 1 km around the station is the origin/destination area. The reason for choosing 1 km is that the comfortable pedestrian zone of the residents is 500 m (Tian Zongxing, 2018), considering that passengers can reach the station by other transport mode, like bike or bus, the comfort distance of residents to the subway station can be increased to 1 km.

After obtaining the origin/destination area of each passenger, the spatial distribution of each group of passengers can be known by combining the geographic data. And the distribution will be be expressed as a percentage of passengers observed.

It should be noted that the origin/destination area obtained may not be the actual origin/destination of the passengers, especially for the SC data where only one mode of transportation is included. For example, only the SC data of the subway transaction record. The final identified origin/destination area will be located near each subway station. However, the passenger's travel may involve the transfer of different modes of transportation, so the subway station used to transfer to the bus may be seen as the destination area. In other words, the origin/destination area may not only be a place where passengers often have activities, but also a place to transfer to other modes of transportation.

3.5 LONGITUDINAL ANALYSIS

After passenger characterization, the characteristics of passengers with different trip frequencies at different period are quantified. The quantified results are: daily ridership, passenger proportion in each time interval based departure time (for first and last trip separately, for working day and non-working day separately), passenger proportion in each distance interval trip distance (for working day and non-working day separately) and passenger proportion near each station based on origin/destination area (only destination locations discuss on working day and non-working day separately). Based on these quantitative results, the evolution of PTCs will be obtained by longitudinal analysis. The benefits of using longitudinal analysis to explore PTCs evolution are mentioned in the study of Chu (2015), that is allowed to measure the change as individual level. And relevant information of each passenger can be found from SC data. Therefore, longitudinal analysis is the method applicable to this study.

Longitudinal analysis requires the collection of information on the same set of variables produced by the same sample members over time Tourangeau, Zimowski, and Ghadialy (1997), and further analysis of this information to understand its changes. Here, the fixed sample members are obtained through data preprocessing. Then, in the passenger classification and passenger characterization, the characteristics of each passenger are quantified, and the quantized results are grouped and displayed according to the different trip frequencies of the passengers. This can not only understand the differences between passengers, but also compare the characteristics of different periods. In understanding the changes in characteristics, the specific method adopted is observation and statistical analysis. The methods are described below.

- **Observation**
  Here, this thesis will first visualize all the quantified results. The daily ridership will present a daily ridership of passengers composed of different travel frequencies

over different periods. The average departure time corresponds to the percentage of passengers distributed during each time interval. The trip distance shows the percentage of passengers distributed over each distance interval. For these three indicators, their distribution shape will be observed to determine whether there is a change. The origin/destination area shows the percentage of passengers distributed in different areas. It will combine the geographic data to show the spatial distribution of the number of passengers through the heat map, and judge whether there is any change through the color depth.

The method of observation is taken because the characteristics of the passenger can be visualized, and the researcher can quickly discover the change of the feature by the difference in shape or color.

- **Statistical analysis**
  Observed results are subjective judgments from the investigator and there may be errors. The two-sample Kolmogorov–Smirnov test (KS test) will be used to analyze the observations, which is a nonparametric test and can be used to test whether two data samples come from the same continuous distribution (Hodges, 1958). The reason for choosing two-sample KS test is that as an non-parametric test, it does not need to know the specific distribution of the data, and has a wider scope of application. The other test method like paired t-test, although it can also be used to determine whether two samples are equally distributed, the premise is that the two samples need to follow a normal distribution. In this study, it is necessary to judge the distribution of the four indicators separately, so the two-sample KS test is selected due to its wider scope of application.

  The null hypothesis of this statistic method is that the samples are drawn from the same distribution. In order to test the null hypothesis by two-sample KS test, the cumulative distribution functions are formed from the two data samples, and then the KS statistic is obtained by computing the maximum difference of the cumulative distribution functions. So, KS statistic is:

$$D = \sup_{x} |F_m(x) - G_n(X)| \tag{3.1}$$

where $F_m(x)$ and $G_n(X)$ are the cumulative distribution functions of the two sample respectively, sup is the supremum function. The null hypothesis is rejected at level α if:

$$D > c(\alpha)\sqrt{\frac{n+m}{nm}} \tag{3.2}$$

Where n and m are the size of the two sample data respectively, the value of $c(\alpha)$ is given in the table for the most levels of alpha. By Equation 3.2, the probability value (p-value) that the null hypothesis is true can be obtained.

In this thesis, the two-sample KS test is conducted by SciPy, which is a Python-based software for scientific computing (Jones, Oliphant, Peterson, et al., 2001–). In SciPy, the KS test follows the study of Hodges (1958), and returns two values, that are a KS statistic and a two-tailed p-value. If the p-value is greater than 0.05, the null hypothesis cannot be rejected, namely the two samples come from the same distribution.

By comparing the distribution of daily ridership and the distribution of passenger proportion based on departure time, trip distance and origin/destination area by KS test, it can be known whether the PTCs in different periods are the same (the same distribution means that the PTCs have no change).

The two-sample KS test can only explain the distribution of two data samples in a statistical sense. That is to say, when most of the data in the two samples are similar, and only a very small number of data are different, the results of the KS test may show that the two samples obey the same distribution. In this thesis, such a situation may cause important changes in a certain day, a certain period of time or local areas to be ignored. Therefore, after the completion of the KS test, obvious changes will be explained separately based on the observed results.

- **Member switching analysis**
  In order to understand the difference in PTCs between passengers, the passengers are grouped according to the trip frequency of each period (see in Section 3.3), and the PTCs reflected in each group are discussed separately. However, passenger's trip frequency will change. That is to say, members of the group with the same trip frequency may be different at different period. In this case, even if the group of the same trip frequency exhibits the same PTCs at different period, these PTCs are represented by different passengers.

  Therefore, member switching analysis is performed after understanding whether the PTCs of each group have changed. The specific method is to explore the changes in the group that each passenger belonging to in different periods. This can help to understand whether changes in PTCs reflected by groups of the same trip frequency are related to changes in members of the group.

3.6 LIMITATIONS AND ASSUMPTIONS

This section will summarize all the assumptions in the methodology and the impact of these assumptions, and explain the limitations of the method itself.

- *Assumption 1: One card only corresponds to one passenger.*
  In reality, there is no guarantee that a card will always be used by the same person, especially an anonymous card, and it is not possible to verify that the users of each card are the same. Eventually, the resulting characteristics may not be from the same passenger. But this assumption is the basis of the methodology. What can be done in this study is to minimize the impact of this hypothesis on the results. The

effect of this hypothesis on the result can be reduced to some extent by calculating the most commonly used departure time, mode of trip distance, and determining the origin/destination area through the most commonly used departure station. Because this can reduce the impact of irregular changes, and if the card is borrowed by others, the resulting different travel characteristics can be ignored as irregular changes. In addition, when selecting sample data, the data generated by the recyclable one-way ticket will not be considered because their users are not the same person.

- *Assumption 2: Characteristics are different if passengers have different trip frequencies.*
  This assumption is mainly for grouping passengers and looking for difference between them. Therefore, after obtaining the PTCs quantization results of passengers with different trip frequencies in passenger characterization, the assumption can be verified by performing comparison during the same period.

- *Assumption 3: Passengers always choose the shortest path.*
  In reality, passengers have many other factors to consider when choosing a route, in addition to the length of the route, such as price, number of transfers and congestion. This means that passengers will have different trip distances even if the starting and ending points are the same. This is an inevitable problem in SC data that cannot track path selection. However, the shortest path often means less travel time, and most people tend to have shorter travel times, and the impact of this assumption is limited. Therefore, the impact of this assumption will be ignored in subsequent chapters.

Some limitations and related explanations of the methodology are described below.

- *Limitation 1: Did not consider all passengers in SC data.*
  The data samples selected for the longitudinal analysis are only part of the raw data, so the resulting PTCs do not represent the original data. However, the composition of the passengers that produce the raw data is not the same as the composition of the passengers in the sample data, so the characteristics of their performance should be different.

- *Limitation 2: The travel mode involved in SC data is not comprehensive.*
  The SC data will only include the modes of transportation that smart cards can use, such as subways, buses and intercity railways. Therefore, in this thesis, the terminology used is only for the modes of transportation included in the data. For example, when using smart card data for the subway, the trip generated by the passenger is specifically generated by using the subway. This limitation is brought about by the data itself. As with the Limitation 1, the passenger characteristics involved in different modes of transportation should be different. When analyzing based on SC data, the PTCs obtained are only for the transportation methods involved in the data.

- *Limitation 3: For departure time, passengers with only one trip per day have no difference between departure time of first and last trip.*
  The impact of Limitation 3 is mainly due to the proportion of only one trip a day. If this proportion is high, there will be no difference in the departure time between the

first travel and the last trip. Therefore, the difference of the departure time for first trip and last trip can represent how large the proportion.

- *Limitation 4: When calculating departure time, trip distance and origin/destination area, if there is a result with the same frequency, only one of them can be selected.*
  The calculation of these three indicators is based on the results obtained by the frequency, which is mainly to find the most representative PTCs and avoid the effects of irregular changes. If the average departure time or average trip distance is calculated, it may be affected by extreme values caused by irregular change. But it may happen that the frequency is the same. For example, of the ten trips associated with a passenger, five trips are made at 8:00 and five times at 9:00. When such a situation occurs, the passenger's departure time is determined based on the earliest relevant trip. The consequence of this is that some of the representative characteristics may be ignored, but the focus of this study is not on describing PTCs as completely as possible, so no further discussion is done.

4 RESEARCH BACKGROUND AND DATA INTRODUCTION

The data used in this study is the SC data of Shenzhen Metro in December 2012, December 2013 and December 2014. This chapter introduces the city situation in Shenzhen, the network and operation of the Shenzhen Metro, and a detailed description of the information contained in the SC data.

4.1 SHENZHEN CITY

Shenzhen, located on the east bank of the Pearl River and across the water from Hong Kong, is a coastal city in southern China (Figure 4.1). As the first economic zone which was established in 1980, Shenzhen's urban construction and economic development have been greatly supported by the Chinese government. After becoming a special economic zone, the population of various provinces moved to Shenzhen, which made Shenzhen an immigrant city and experienced explosive population growth. Until 2017, the resident population has increased from more than three hundred thousand to twelve million. Meanwhile, Shenzhen's GDP rose from 197 million yuan in 1979 to 2.24 trillion yuan in 2017 (Shenzhen municipal people's government, n.d.). In order to meet the development needs of the city and the increasing travel demand of residents, Shenzhen's public transportation system has also experienced rapid development. By 2016, the scale of rail transit operations in Shenzhen reached 178 kilometers. As for bus system, 883 kilometers of bus lanes and 2,903 bus station were built, and 909 bus lines were put into operation. At peak times, public transport travel accounted for 56.1% of motorized travel. In addition, public bicycles have been launched to enrich the travel patterns of residents (Shenzhen municipal transportation commission, 2016).

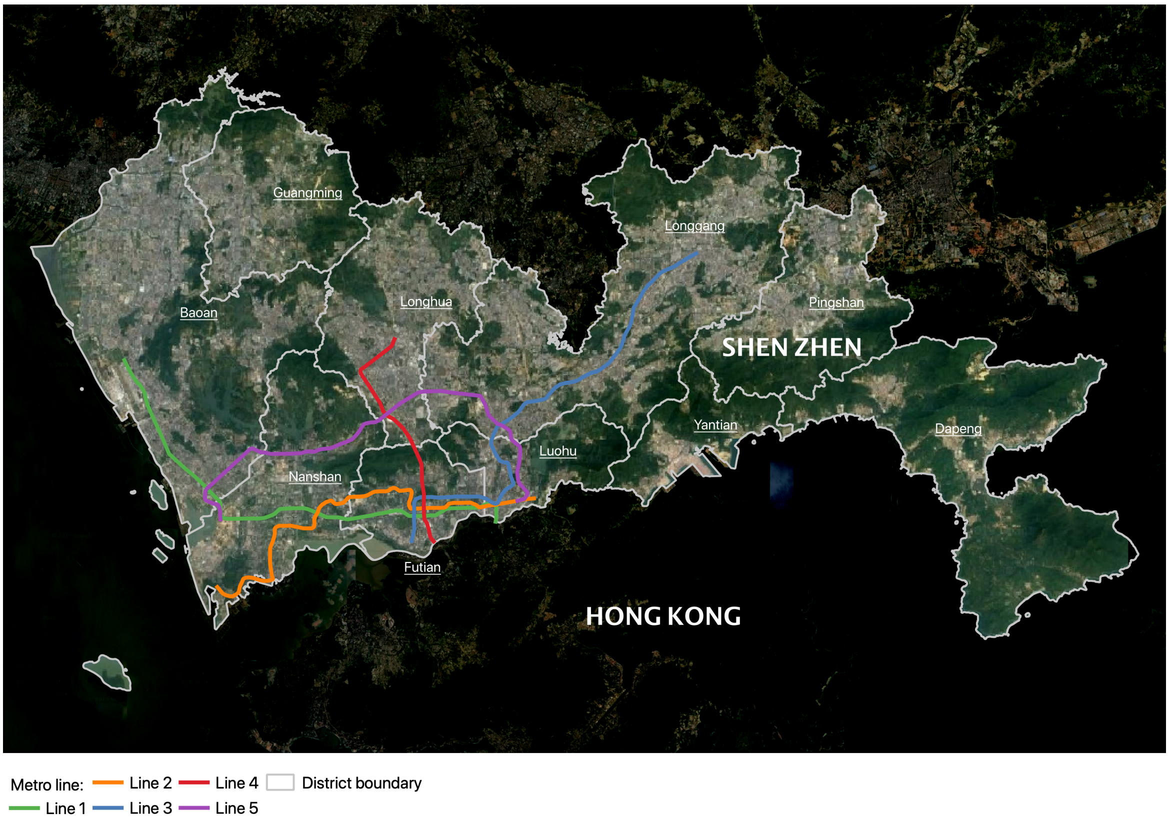


Figure 4.1: Map of Shenzhen in 2012-2014

In this context, this thesis will explore the evolution of PTCs of Shenzhen residents when they use the subway.

4.2 SHENZHEN METRO

The Shenzhen Metro was first opened in 2004. Based on the time scale contained in SC data, this section details the network, fare policy and ticket type of Shenzhen Metro during 2012-2014.

4.2.1 Network

From 2012 to 2014, the network of Shenzhen Metro consisted by 5 operational metro lines with 118 stations, without any new lines and stations. The network connects three railway stations (by Shenzhen North station, Buji station, and Luohu station, represented by a blue circle in Figure 4.2) and one airport (by Airport East station, represented by a red circle in Figure 4.2). In the three years, the annual ridership was 718 million, 917 million and 1036 million respectively. In the case that the network has not changed, the Shenzhen Metro has adopted a solution to increase the number of trains operated daily, the relevant data is listed in Table 4.1.

Table 4.1: Shenzhen metro operation summary in 2012-2014 ("2013 annual statistical analysis report of urban rail transit in China", 2014; "2014 annual statistical analysis report of urban rail transit in China", 2015)

Year	*Annual ridership (million)*	*Average number of operation trains per day*
2012	718	-
2013	917	1,997
2014	1,036	2,120

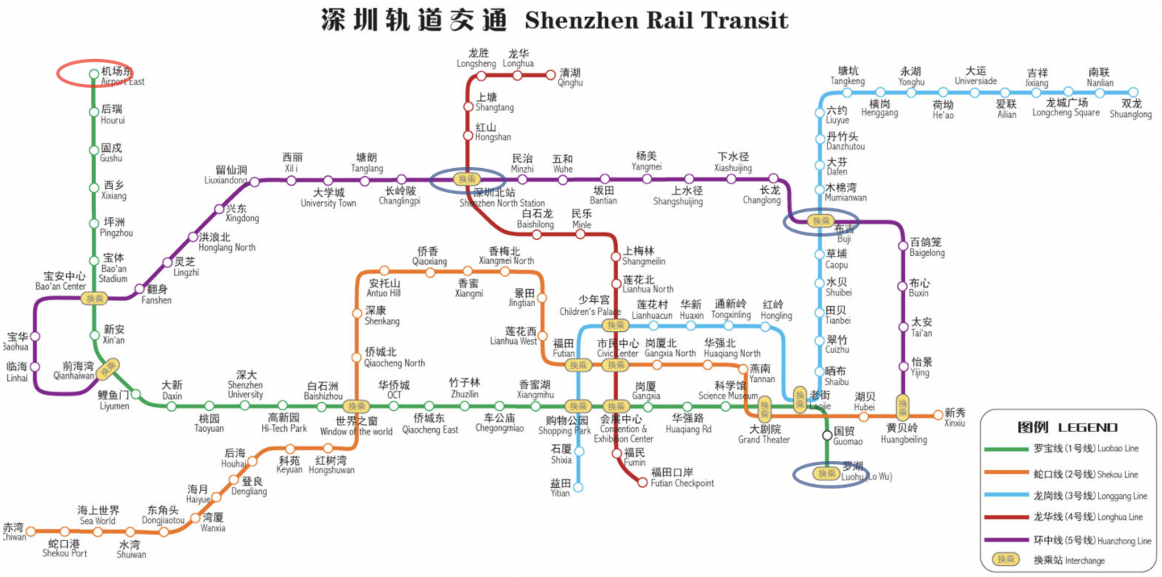


Figure 4.2: Shenzhen Metro network

Other details of the network would be introduced below through each metro line. The map of network is shown in Figure 4.2

- **Line 1**
  Line 1 (the green line) is the earliest metro line commenced by Shenzhen Metro, which is divided into three sections to complete the construction at different times. The first section from Luohu to Window of the World commenced on 2004, and the other two section commenced on 2009 and 2011. Line 1 is the busiest line in Shenzhen Metro, which used to connect the Shenzhen railway station, Shenzhen Baoan International Airport, Shenzhen central district and some important business districts. Line 1 runs westward from Luohu which to Airport East, and is managed by SZMC (Shenzhen Metro Group). And passenger can transfer to Hong Kong Metro at Luohu.

- **Line 2**
  Line 2 (the orange line) is the second commenced line in Shenzhen, which is also generally east-west as Line 1. The section from Central Theater to Window of Word is parallel to Line 1, this is because Line 2 has an important function to release the

transportation pressure of Line 1. However, most of the areas passing through the west line of the Window of the World in Line 2 are areas to be developed, so there are fewer passengers on the Line 2 west. SZMC is the operator of Line 2.

- **Line 3**
  Line 3 (the blue line) connects the south and east of Shenzhen, which commenced on 2010. A subsidiary of SZMC, named Shenzhen Metro No.3 Line Operations is responsible for the management of Line 3. Line 3 is the only urban rail transit connecting the northeast region of Shenzhen with the city center. It also connects the three administrative districts of Futian, Luohu and Longgang, strengthening the links between the regions. Therefore, Line 3 bears a considerable part of the passenger flow pressure between these regions.

- **Line 4**
  Line 4 (red line) runs south-north from Futian Checkpoint to Qinghu, connecting Futian and Longhua District of Shenzhen. After phase 2 of Line 4 is completed and put into operation in 2010, the right to operate Line 4 is owned by MTR Corporations (Shenzhen) which is the subsidiary of MTR. Futian Checkpoint is another station where can transfer to the Hong Kong metro in addition to Luohu station on line 1. In January 2014, due to the serious congestion during rush hours, Line 4 was planned to gradually add all four marshalling trains to six, which lasted until January 2015.

- **Line 5**
  Line 5 (the purple line), operated by SZMC, opened in 2011. Line 5 is a semi-circular line, starting from Qianhaiwan in the west and ending at Huangbeiling in the east. It mainly covers the surrounding area of the central district of Shenzhen.

All the information above comes from Shenzhen Metro Group Co., Ltd. (n.d.); Shenzhen municipal people's government (n.d.); Southern Urban Daily (2014). And some other details are shown in Table 4.2.

Table 4.2: Station, length and operator of Line 1-5

Network	*Station*[1]	*Length (KM)*	*Operator*
Line 1	30	41.04	SZMC
Line 2	29	35.78	SZMC
Line 3	30	41.7	Line 3 Operations
Line 4	15	20.5	MTR Corporation (Shenzhen)
Line 5	27	40.0	SZMC

[1] Station includes transfer station

4.2.2 Fare policy

The metro line is run by several companies, but all the companies use the same fare policy, and passengers can transfer between lines without leaving the station. This section will introduce the basic fare pricing method and fare discount scheme in the fare policy (Shenzhen Metro Group Co., Ltd., n.d.).

The basic fare pricing method will be calculated in stages according to the mileage of the ride. In the stage where the mileage is within 4 kilometers, only 2 yuan is required; for the stage of 4 to 12 kilometers, every 1 yuan can take 4 kilometers; for 12-24 kilometers, every 1 yuan can take 6 kilometers; more than 24 kilometers, every 1 yuan take 8 kilometers. For passenger, the advantage of the fare policy is that the farther they travel, the less money they spend per kilometer.

In the discount scheme, senior citizens over 65 years old can show their IDs to the staff and take the subway free of charge. In addition, active military personnel, retired cadres, and disabled people can also take the subway free of charge with relevant documents. In addition to the relevant documents, persons with disabilities who have a household registration in Shenzhen can apply for a special preferential card. The difference between the use of the certificate and the card is that the relevant travel data will be recorded only when the ride card is used, and there is no data record when the certificate is used.

For children, the discount scheme varies according to the age and height of the child. Children under 1.2 metres or under 6 years of age can travel free of charge with their ID or child ride card. Children between 1.2-1.5 metres or between the ages of 6-14 enjoy a 50% discount on the fare and use a discounted one-way token with the voucher or a special preferential card for child to travel.

For students, primary and secondary school students under the age of 18 can apply a special preferential card for student and enjoy a 50% discount on the fare, so their characteristics can be tracked based on the data generated by the card.

The above discounts are for some special groups. In addition, in order to promote Shenzhen Tong, a smart card, subway passengers who use it can enjoy a 5% discount of the ticket price. And Shenzhen Tong can also be used to take the bus, passengers can use it to transfer between the subway and the bus while get a discount of 0.2 yuan.

4.2.3 Ticket type

Depend on the issuer, tickets used in Shenzhen Metro can be divided into three types (Shenzhen Metro Group Co., Ltd., n.d.):

- **Identification**
  The identifications are mainly used by special people who can enjoy the free ride discount mentioned in the previous section. The essence of the identification is not a ticket card, but because it can be used as a voucher to allow eligible passengers to take the subway, it has the function of a ticket card. The issuer of the identification is the government department.

  The specific method of use is that the passengers present the relevant identifications to the staff to prove their special status. After checking and confirming the validity, the staff will open a dedicated passage, and the passengers enter and take the subway through the dedicated passage. When leaving the station, it is also necessary to exit the station through a dedicated channel after the staff has checked the identifications. Because the verification of the identifications and the management of the dedicated passage are all done manually, the relevant travel data cannot be recorded into the database.

- **Shenzhen Tong**
  Shenzhen Tong issued by Shenzhen Tong Limited, is a reusable contactless stored value smart card which can be used for electronic payment in public transport system (Shenzhen Tong Limited, 2013). Shenzhen Tong first needs to apply at the designated location and deliver the deposit to Shenzhen Tong Limited.

  Ordinary passengers who use Shenzhen Tong can not only enjoy a discount of the subway fare, and they can also enjoy certain discounts when they take the bus and transfer between different bus or subway routes. These discount schemes are designed to encourage people to use Shenzhen Tong. Special passengers including disabled, children and students can apply for a courtesy card at the Shenzhen Tong location and enjoy the relevant discount mentioned in the previous section.

  The specific use method of Shenzhen Tong is that the passengers deposit the money into card in advance, and then use the card to check in at the card reader on the gate, and then the gate opens the passage, the passenger enters the station; the same when exiting the station. Each Shenzhen Tong has a unique ID, and the passenger's travel information is recorded immediately when the passenger checking in/out by the card. The search by ID can obtain the travel data of the designated passengers over a period of time, and the passenger characteristics can be identified based on the data. However, except for student cards and courtesy cards, all Shenzhen Tong are anonymous cards and may be used by different users.

- **RFID tokens**
  The Shenzhen Metro has issued an RFID token as a one-way ticket for the subway system. The use of this token is similar to that of Shenzhen Tong. The main difference is that Shenzhen Tong can be used repeatedly. The token can only be used once and will be recycled when it leaves the station. There are no discounts on using tokens.

  Tokens need to be purchased at a ticket vending machine or manual ticket office in the station, and there is a using time limit. Therefore, compared with Shenzhen Tong, the convenience of tokens is lower.

  Each token also has a unique ID and records the user's travel data. However, because the token will be recycled after each use, it means that the passengers used each time are different, it is impossible to obtain multiple sets of data of the same passenger by finding the ID of the token.

The above are the main types of ticket cards as of 2012-2014. Other types of tickets, such as commemorative tickets, are not included in the scope of this study because of the small number of releases.

4.2.4 others

Because the arrival time of the first train and the last train is different in every station, the operation time of each stations is also different. In order to facilitate the calculation, a unified operating time will be used in this study, that is 06:00-24:00. 06:00 is determined according to the earliest operation time of Luohu Station in all stations, and 24:00 comes from the latest end of operation time. Therefore, 06:00-24:00 can cover the operating hours of all stations, avoiding the situation that transaction data is mistakenly ignored due to different operating hours.

4.3 SC DATA AND NETWORK DATA

4.3.1 SC data

The raw data includes the relevant records of all passengers using Shenzhen Metro in December 2012, December 2013 and December 2014 for a total of three months. A record is generated each time a passenger uses Shenzhen Pass or Token to enter or exit the station. The amount of records contained in the raw data and the number of cards involved are shown in Table 4.3.

Table 4.3: The number of records and cards in raw data

Year	*Number of records*	*Number of cards*
2012.12	100,229,525	6,513,259
2013.12	115,143,474	6,412,110
2014.12	127,656,462	6,997,706

It can be seen from Table 4.3 that the change in the number of cards is fluctuating. The number of records related to travel has increased year by year, in line with the growth of the ridership mentioned in Table 4.1. The types of data contained in each record are listed in Table 4.4.

Table 4.4: Data type

Card ID	*Card type*[1]	*Transaction tyepe*	*Station*	*Gate machine*[1]	*Transaction time*
987727139	98	21	1268012000	268012203	″2013-11-30,15:52:05″

[1] There is no card type and gate machine in the December 2012 record

Not every type of data contained in the record will be used in this study, where the gate machine is not needed. Next, the types of data that will be used for analysis in this study are introduced.

- **Card ID**
  All smart cards that can be used to take the subway have a unique ID, which can be used to track the records generated by the same card. In raw data, total number of cards that at least have one record is 14,256,748.

- **Card type**
  The smart cards that can be used to take the subway include ordinary Shenzhen Tong (common card), Shenzhen Tong student card (student card), Shenzhen Tong discount card for the disable (courtesy card), one-way ticket (tokens) and employee card used by subway employees. These cards are represented by different numbers depending on the type. In the raw data, the distribution of card type is shown in Figure 4.3:

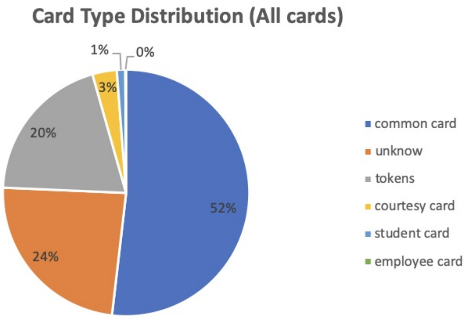


Figure 4.3: The card type distribution of all recorded cards, the unknown card is because there is no record of card type in 2012

As can be seen from the distribution in the figure, most of the cards are common Shenzhen Tong (common cards), and the token accounts for only 20% of the total. In other words, the passengers of the Shenzhen Metro are still mainly based on Shenzhen Tong.

- **Transaction type**
  The transaction type describes whether the smart card is used for check-in or check-out, the check-in is represented by the number '21', and the check-out is '22'. Therefore, the number of records in Table 4.3 actually includes records of check-in and check-out.

- **Station**
  The data in the station column records the station where the passenger check-in/out. With the combination of the previous transaction type, it can also determine whether the station where the passenger is located is the starting point or destination of his trip. This type of data can help understand the spatial pattern of passenger travel behavior. As mentioned in the background study, the number of stations in 2012-2014 is 118, these stations have their own unique ID when recording.

- **Transaction time**
  The transaction time record the time when the passengers check-in/out. The difference between check-in and check -out can be judged by the transaction type. The use of SC avoids two consecutive check-in records and two consecutive check-out records. In the absence of missing data, the check-out and check-in records alternate. And in a trip, always the check-in record is generated before the check-out record. Therefore, after all the records of a card are sorted by transaction time, the check-in

and check-out records for each trip can be found. Transaction time is critical when describing PTCs from temporal perspective.

4.3.2 Network data

The network data contains the name, ID and latitude and longitude of each station, as well as the number, geographical location and length of each line.

4.4 LIMITATIONS IN SC DATA

After understanding the data and background, summarize the limitations they have caused.

- *Limitation 1: SC data only contains travel data for subway passengers.*
  As mentioned in the introduction of Shenzhen Tong, Shenzhen Tong can be used to take the subway and bus, and can enjoy certain discounts. However, the SC data used in this study only contains records when taking the subway, so the final PTCs described are only for subway passengers. In addition, this may cause some errors. For example, the destination area found according to the SC data only for subway may not be the final place of activity of the passenger, but the place to transfer to the bus. Because there is no relevant SC data for bus, the what impact of such errors may cause cannot be discussed, and the error is not within the scope of the study. In subsequent chapters, the errors caused by this limitation will be ignored.

- *Limitation 2: SC data contains a large number of anonymous and one-way tickets, while passengers using the identifications are not recorded.*
  The people who can use the relevant identifications to take the subway have already been introduced. They have different social population characteristics from those who use Shenzhen Tong and one-way tickets, so PTCs should be different, and no further discussion is made. According to the methodology and the introduction of the one-way ticket for the Shenzhen Metro, the records related to the one-way ticket will be deleted here. As for the anonymous cards, their impact also has been discussed in the methodology.

- *Limitation 3: Only one month of data per year in SC data*
  The existing data contains data only for December in different years. Therefore, when exploring the evolution of PTCs, only the changes of day-to-day and week-to-week within one month or year-to-year in different years can be discussed. Here, this thesis chose to explore evolution based on PTCs found in travel data for the same month in different years. However, the PTCs reflected in one month's data are different from the one-year data. For example, one year's data can reflect seasonal changes, but one month's data is not. This thesis chose to ignore such differences, because in China, there is no statutory holiday in December, and it is not a holiday time for primary and secondary school students, so the December data is stable and representative.

- *Limitation 4: Not sure about other factors related to passenger travel.*
  The factors mentioned in this limitation include non-human factors such as weather conditions, train failures and other emergencies. Because the time involved in SC data is early, it is difficult to collect data on other factors, and the focus of this study is not to discuss the impact of external factors on passenger behavior. Therefore, it is assumed here that the non-human factors in the study periods are the same and unchange.

5 RESULT

This chapter details the analysis result of SC data from Shenzhen Metro. All the analysis result is presented by four indicators seperately.

5.1 SAMPLE DATA

The sample data is determined according to the method in data preprocessing and the SC data from Shenzhen Metro. In the sample data, only records for four consecutive weeks in each month are retained, in order to ensure the comparability of the data in time for longitudinal analysis. Therefore, the analysis timeframe consists of three periods, and the daily operation time is 06:00-24:00 during each period. The three periods are:

- Period 1: December 3, 2012 to December 29, 2012
- Period 2: December 2, 2013 to December 28, 2013
- Period 3: December 1, 2014 to December 27, 2014

After deleting the records outside the analysis timeframe and the incomplete and erroneous records, only Shenzhen Tong, which has transaction records in all three periods, is retained. These Shenzhen Tong transaction records constitute the sample data used in this thesis. The number of Shenzhen Tong card involved in the sample data is 1,025,052, accounting for 7.2% of all cards, including the common card, student card and courtesy card. The distribution of the three different types of Shenzhen Tong is shown in Figure 5.1.

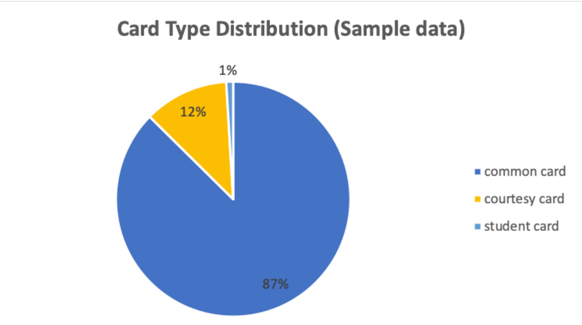


Figure 5.1: The card type distribution in sample data

As can be seen from Figure 5.1, most of the cards are common cards. The courtesy card accounts for 12%, while the number of student cards is further smaller. The reason for the small number of student cards may be that the Shenzhen government stipulates that primary and secondary school students must enroll in the school closest to their home, so that most students only need to walk to the school, namely they do not need to take the subway to go to school. In general, it is not appropriate to classify passengers by card type. The main reason is that the common card contains most of the passengers, and classification according to the type of card will ignore the difference between these passengers.

Among the sample data, 1,025,052 Shenzhen Tong produced different numbers of records in different periods. Table 5.1 shows the number of these records and the proportion of those records in the raw data. The Table 5.1 shows that the number of records involved in sample data has a small range of fluctuations within three years. At the same time, their proportion in the total annual record is decreasing year by year. Combined with Table 4.3, the total number of records in raw data is increasing year by year, which shows that the increase in the number of transactions in the three years is not mainly due to the passengers involved in the sample data. Moreover, although the percentage of cards in sample data is only 7.2% of the total cards, the number of records generated is kept at about 20% of the total number of records, which proves that the passengers of these cards are more active subway users.

Table 5.1: Transaction records in sample data

Period	*Number of records*	*Percentage of total records*
Period 1	22,382,940	22%
Period 2	24,318,703	21%
Period 3	23,390,407	18%

5.2 PASSENGER CLASSIFICATION

In Section 3.3, why trip frequency is used to classify passenger has been introduced. With the card type distribution, it also explains why card type is not a good choice here to classify passenger. Next, this study details the trip frequency of passengers in different periods by using the transaction data of each card in the sample data. Here, one trip contains one check-in and one check-out record.

Figure 5.2 shows the distribution of passenger trip frequencies over different periods. The figure selects the trip frequency interval to be displayed for 5. The smaller the interval, the more detail that can be displayed. But because the frequency distribution here is wide (0-200), and most of the passengers are concentrated in the range where the travel frequency is less than 60, if the interval is too small, no obvious trend can be observed, and too much spacing (like 10) will ignore the details of the concentrated area of the passenger.

It can be seen from the figure that the distribution of trip frequency is similar in the three periods, and the number of passengers in each interval within three periods is not much different. In sample data, about 500,000 passengers travel at a frequency of 1-5 times

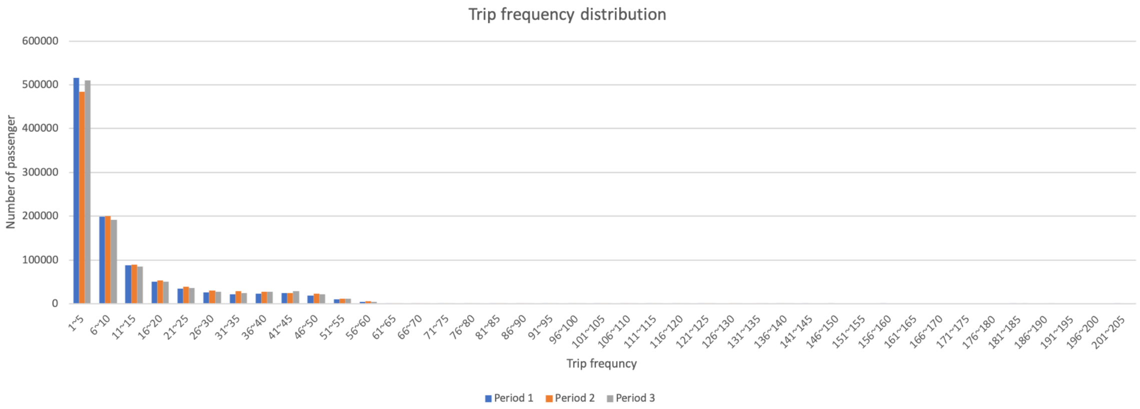


Figure 5.2: Trip frequency distribution during three periods in analysis timeframe

a month, accounting for half of the total number of people. As the trip frequency increases, the number of passengers in the corresponding interval decreases exponentially.

Figure 5.2 also shows that the difference in trip frequency is large (from 1 time per month to 205 times per month) and the passenger number also varies greatly (the largest is 500,000 and the least is only 1). In order to catch the difference among all passengers as much as possible, considering the characteristic of passenger's trip frequency distribution mentioned above, the trip frequency in Figure 5.2 is divided into three intervals for the passenger classification. The details about the three interval and why they so divided are as follows:

- **Trip frequency in 1-5 (TF 1):** Although the frequency of their use of the subway is low, but the number of passengers is large, the total number of trips eventually generated is also considerable. Therefore, the characteristics of this part of the passengers deserve to be studied separately.
- **Trip frequency in 6-28 (TF 2):** Figure 5.2 shows that half of the passengers in the sample data are distributed in the range of 6-205, which is quite wide. It is reasonable to doubt that these passengers should have significant different PTCs since their trip frequency has a greatly difference. Therefore, the travel frequency is divided into two intervals of 6-28 and 29-205. Among them, passengers with a travel frequency in the range of 6-28 have the average number of trips per day (28 days in each period in analysis timeframe) is less than or equal to one; Passengers in the interval 29-205 have the average number of trips per day is greater than one.
- **Trip frequency in 29-205 (TF 3):** Same as TF 2.

When dividing the interval of the trip frequency, the more the number of divisions, the more comprehensive the difference between passengers. One purpose of this study is to show the difference in the evolution of PTCs between different passengers, not to find a passenger classification method to show the difference as much as possible. Considering the purpose of the study and the calculation time, this study chooses to divide the passengers into three groups. The results of grouping passengers are shown in Table 5.2 based on the three trip frequency intervals divided.

Table 5.2: Passenger classification based on TF 1, TF 2 and TF 3 in three periods

TF 1: Trip frequency in 1-5			
Period	Period 1	Period 2	Period 3
Number of Passengers	515,559	484,898	510,233
Number of trips	1,368,607	1,310,891	1,343,987
Average trip frequency per passenger	2.65	2.70	2.63
TF 3: Trip frequency in 6-28			
Period	Period 1	Period 2	Period 3
Number of Passengers	390,055	401,592	381,682
Number of trips	4,803,878	5,016,618	4,749,470
Average trip frequency per passenger	12.32	12.49	12.44
TF 3: Trip frequency in 29-205			
Period	Period 1	Period 2	Period 3
Number of Passengers	119,438	138,562	133,137
Number of trips	5,027,245	5,842,317	5,601,851
Average trip frequency per passenger	42.09	42.16	42.08

As can be seen from Table 5.2, for the passengers involved in the sample data, the number of passengers distributed in the same travel frequency interval is different at different times. For TF 1 and TF 2, the overall number of passengers has decreased, and only TF 3 has increased. That is, during the analysis timeframe, some passengers increased the frequency of travel by subway in the sample data. This can also be reflected in Table 5.1, the increase in transactions records. However, in TF 1-3, the average number of trips per passenger did not change significantly, so the evolution of PTCs could not be found here. It is necessary to describe in more detail about the PTCs of each trip frequency interval in different periods.

The PTCs of the sample passengers in one period are composed of PTCs represented by the three trip frequency intervals of the same period, for example: PTCs in Period 1 = PTCs in TF 1 in Period 1 + PTCs in TF 2 in Period 1 + PTCs in TF 3 in Period 1. As long as the PTCs in each interval of the same period are obtained, the PTCs of this period can be found.

5.3 PTCS AND DIFFERENCE AMONG PASSENGERS

Here, all the findings about PTCs would be shown by figures, which is used to discuss the difference among passengers.

5.3.1 Daily ridership

From Figure 5.3, the most difference between passengers with different trip frequency is the ridership on working day and non-working day. Passengers in TF 1 generate more trips on non-working, which passengers in TF 3 is the opposite, generating more on working day. There is no difference for passenger in TF 2. Such differences are observed in all three periods.

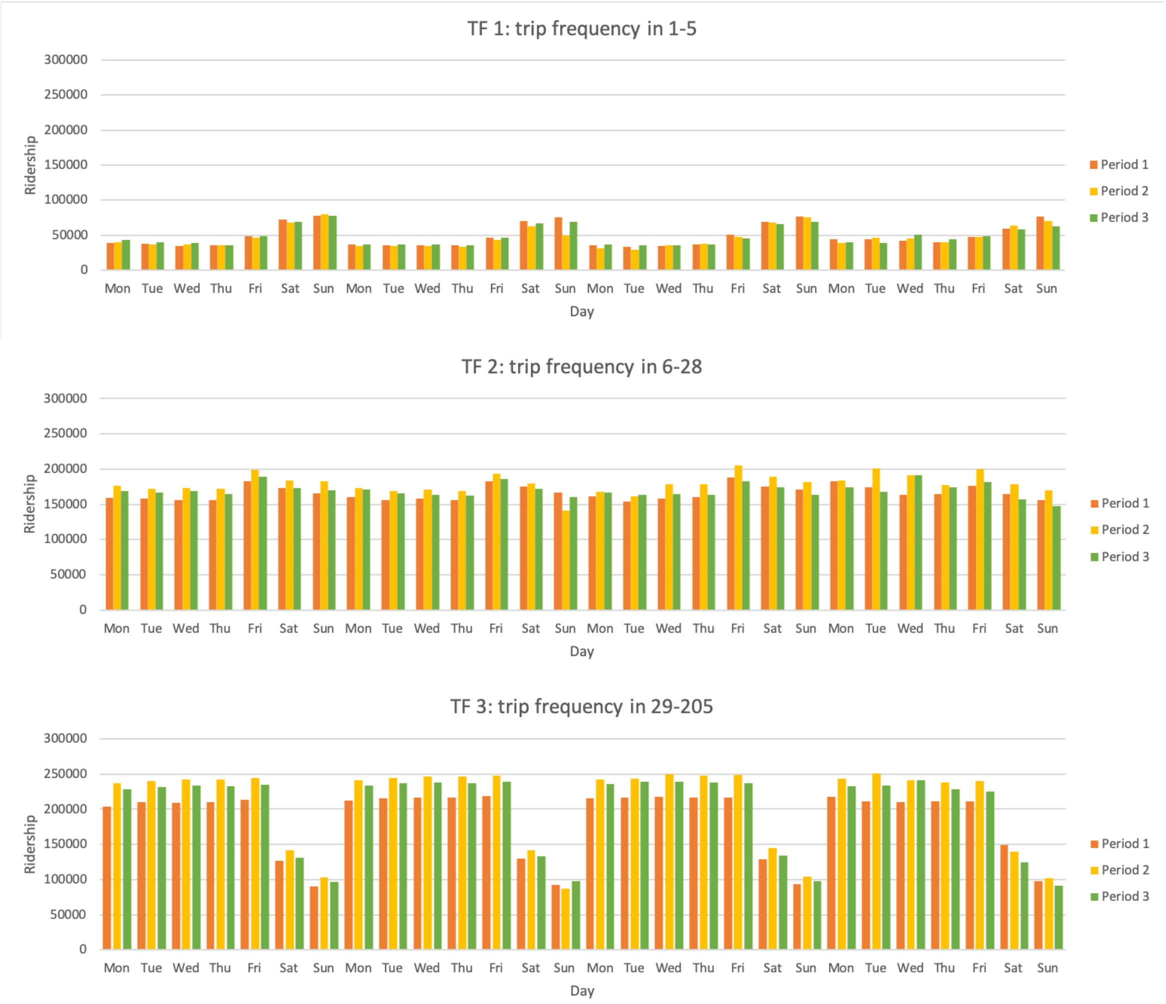


Figure 5.3: Daily ridership for passengers with different trip frequency in the three periods

5.3.2 Departure time

Figure 5.4 shows the most frequent departure time for first trip and last trip on working day and non-working day separately. In every subfigure, the distribution is similar during every period, so the difference among passengers is similar.

Passengers with TF 3 is obviously different from others, their departure time for first trip concentrate on 08:00-08:30 and for last trip is on 18:00-18:30 on working day. On non-working day, passengers with TF 3 also have same central tendency as them on working day, but less obvious.

No matter on working day or non-working day, such central tendency cannot observe in passengers with TF 1 and 2, which means that passengers have an evenly distributed on departure time when their trip frequency in TF 1 and 2. In addition, the similarity for the departure time of first trip and last trip reflects that passengers prefer single trip rather than round or multiple trips per day when their trip frequency in TF 1 and 2.

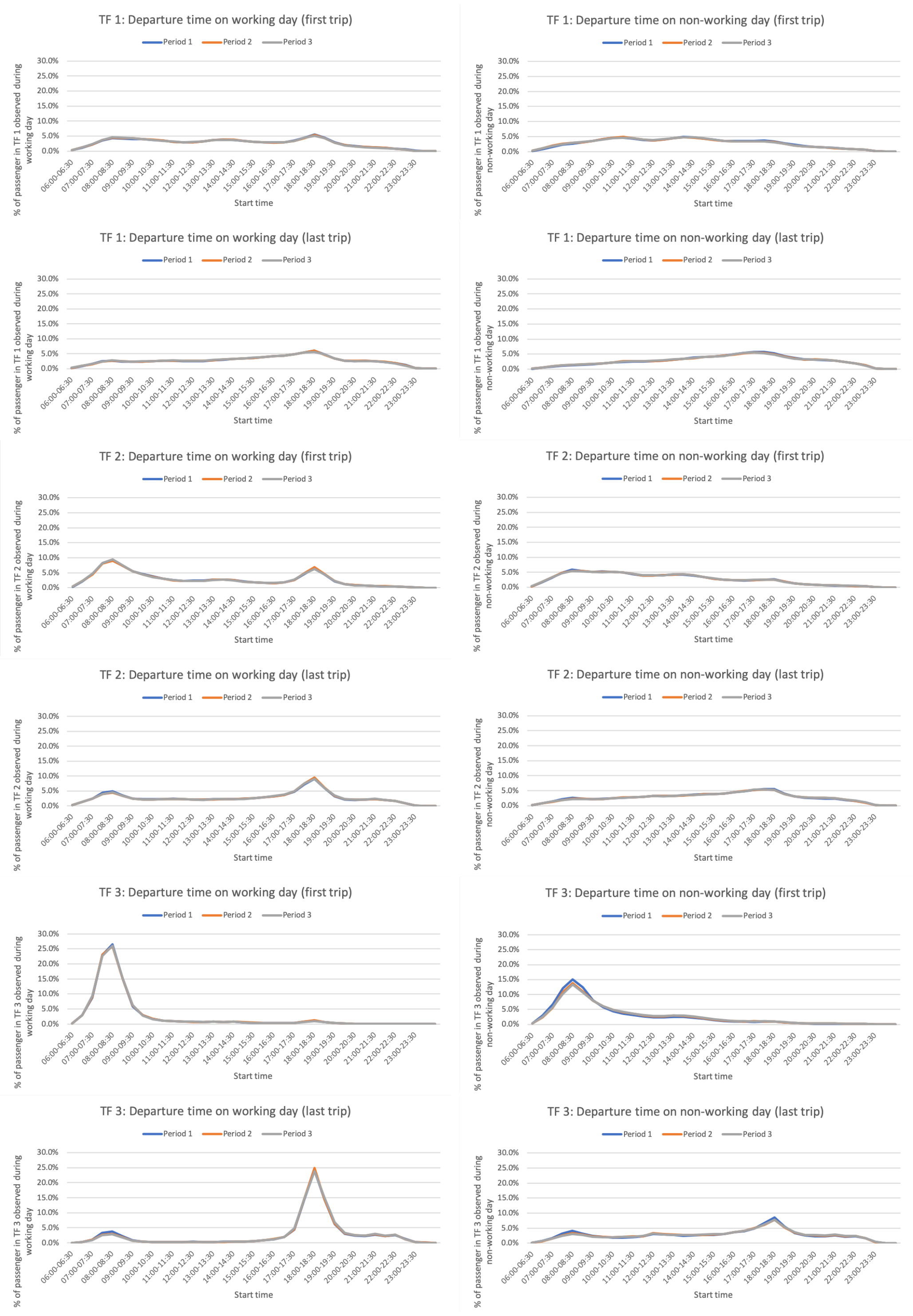


Figure 5.4: Departure time for passengers with different trip frequency in the three periods

5.3.3 Trip distance

Figure 5.5 is the trip distance distribution on working day and non-working day, and during every period, the shape of the distribution in every subfigure is similar. On working day, the difference is reflected in that passengers with TF 3 have a longer trip distance than others, and the proportion of passenger is fluctuated from 3 km to 12 km. The passengers in TF 1 and 2 have trip distance concentrated at 3 km, and the concentration is more pronounced for passengers with TF 2, while near 10% of the passengers is at 3 km. On non-working day, the shape of the distribution is more similar than that on working day. Except for the passengers with TF 1, the distribution is slightly less concentrated. In summary, the difference is more significant on working day than non-working day.

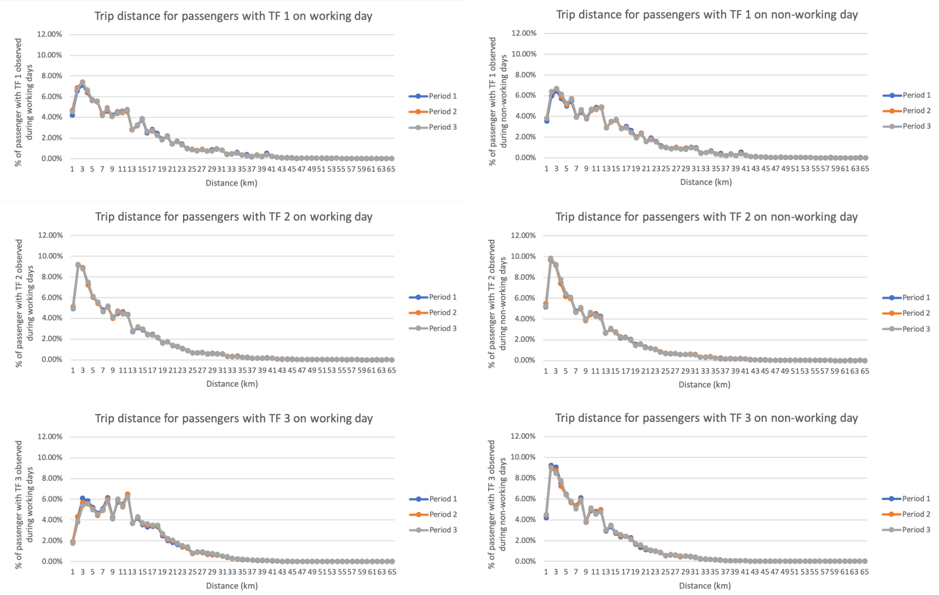


Figure 5.5: Trip distance for passengers with different trip frequency in the three periods

5.3.4 Origin/destination area

Figure 5.6 shows the passenger distribution of origin/destination area on Period 1. For the passengers with a low trip frequency, that is TF 1, the origin/destination area is calculated based on their first trip per day. For the other passengers with relatively high trip frequency, TF 2 and 3, it would base on their first and last trip per day. The specific method is detailed in Chapter methodology.

From Figure 5.6, in Period 1, the passengers with different trip frequency shows the greatest differences in the distribution of origin area. For passengers with TF 1, the origin area with more passengers is near the Luohu station, Laojie station and Huaqiang North station. For passengers with TF 2, the areas with more passengers is similar to that for TF 1, but less concentrated. For passenger with TF 3, the origin area with most passenger is near the Pingzhou station.

The passenger distribution of destination area on working is similar when passengers with TF 1 and 2, which is near the Luohu station, Laojie station and Huaqiang North station. While for passenger with TF 3, destination areas with more passenger are all around Line 1, mainly located near Laojie station, Huaqiang North station, Shopping park station, Chegongmiao station and Shenzhen university station. For destination area on non-working day, the passenger distribution is all concentrated on Luohu station, Laojie station and Huaqiang North station and there is no obvious difference.

The difference, among passenger with different trip frequency in Period 1 discussed above, can also observe in Period 2 (Figure A.1). But in Period 3 (Figure A.2), the difference has changed. For passengers with TF 1, the origin areas with more passengers change to Shenzhen North railway station and Luohu Station, while the distribution for other passengers is similar to before. On working days, the passenger with TF 1change to distribute more near Shenzhen North railway station and Laojie station, which can also be observed on the distribution on non-working day.

In summary, the difference among sample passengers is mainly reflected in the location of area with more passengers.

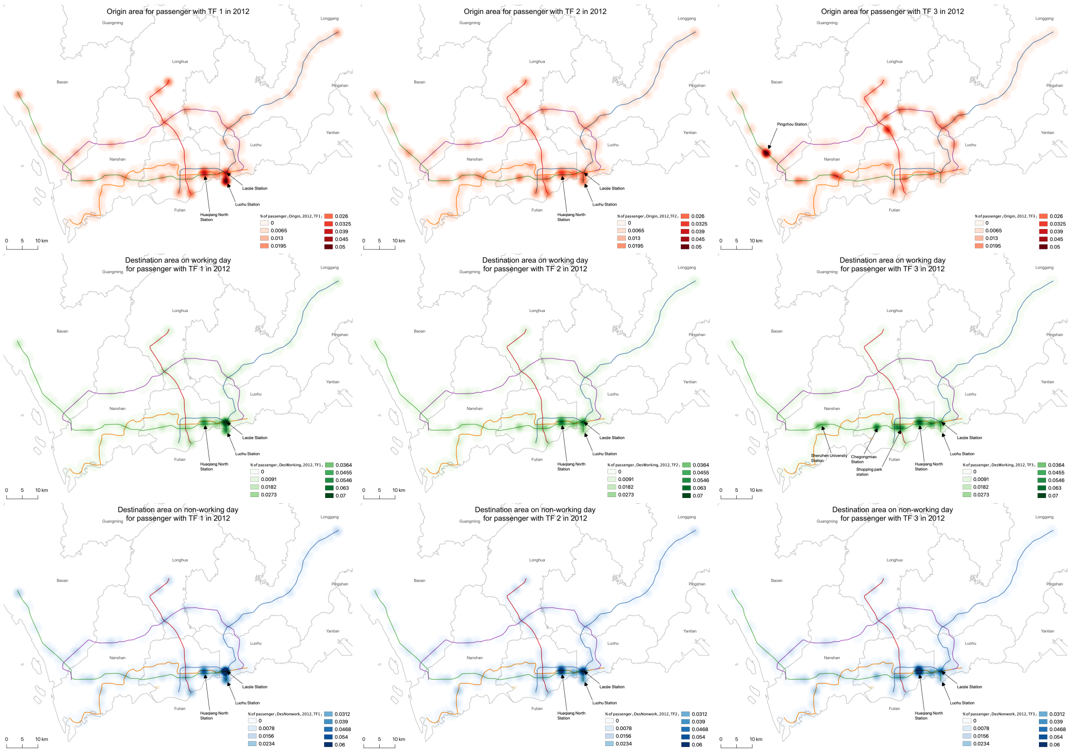

Origin area for passenger with TF 1 in 2012
Origin area for passenger with TF 2 in 2012
Origin area for passenger with TF 3 in 2012
Destination area on working day for passenger with TF 1 in 2012
Destination area on working day for passenger with TF 2 in 2012
Destination area on working day for passenger with TF 3 in 2012
Destination area on non-working day for passenger with TF 1 in 2012
Destination area on non-working day for passenger with TF 2 in 2012
Destination area on non-working day for passenger with TF 3 in 2012
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Laojie Station
Luohu Station
Huaqiang North Station
Pingzhou Station
Shenzhen University Station
Chegongmiao Station
Shopping park station
0 5 10 km
0
0.0065
0.013
0.0195
0.026
0.0325
0.039
0.045
0.05
0
0.0091
0.0182
0.0273
0.0364
0.0455
0.0546
0.063
0.07
0
0.0078
0.0156
0.0234
0.0312
0.039
0.0468
0.054
0.06

Figure 5.6: The origin/destination area for passengers with different trip frequency in Period 1

5.4 STATISTIC ANALYSIS

This section shows the results of a comparison of the PTCs from passengers with different trip frequency by the two-sample KS test. The results of the test mainly include KS statistic and p-value, the smaller the KS statistic, the bigger the p-value. And the null hypothesis is that the two data samples come from the same distribution. In this thesis, an alpha of 0.05 is used as the cutoff for significance, which means if the p-value is less than 0.05, the null hypothesis would be rejected.

5.4.1 Daily ridership

Firstly, the daily ridership generated by passengers with three different trip frequencies in three different periods are summed separately, and then the proportion of daily ridership of passengers with different trip frequency is calculated separately. This is because the number of passengers with the same trip frequency in different periods is different, so the daily ridership of a certain period will be relatively more or less overall, and KS test may mistakenly think that the daily ridership distribution of these two periods is different. But the actual situation may only be caused by an increase in the number of people, and the passenger's travel needs have not changed. The percentage can be used to avoid such errors. Here, the proportion of daily ridership from passengers in each trip frequency is compared. The test results are shown in Table 5.3:

Table 5.3: Two-sample KS test of proportion of daily ridership by passenger with TF 1, 2 and 3 in Period 1, 2 and 3

Period	*Trip frequency*	*KS statistic*	*p-value*
Period 1 vs Period 2	TF 1	0.10714	0.99503
	TF 2	0.14286	0.91681
	TF 3	0.25000	0.30035
Period 2 vs Period 3	TF 1	0.21429	0.49026
	TF 2	0.14286	0.91681
	TF 3	0.14286	0.91681
Period 1 vs Period 3	TF 1	0.28571	0.16875
	TF 2	0.25000	0.30035
	TF 3	0.32143	0.08756

As can be seen from Table 5.3, all p-values are greater than 0.05, which means that the null hypothesis cannot be rejected. This also shows that at a significance level of 0.05, which means the daily ridership proportion by passengers with TF 1/2/3 in the three periods obey the same distribution, and no significant changes occur.

However, as can be seen from Figure 5.3, the daily ridership of Period 2 is generally higher than the other two periods except for the second Sunday (December 15, 2013). In the real world, there are many situations in which abnormal ridership changes, including the occurrence of extreme weather, temporary traffic control, and large-scale gatherings. The sporadic effects of these external factors may cause changes in ridership of one day, but the ridership will also recover after the external conditions return to normal. Therefore, when comparing the distribution of different periods, such an irregular ridership changes of one day should be ignored. In other words, there is no significant change in the proportion of daily ridership of passengers with different travel frequencies.

5.4.2 Departure time

Based on each passenger's departure time of first and last trip, Section 5.3.2 shows the distribution of the percentage of passengers with TF 1, 2 and 3 over on day. Here, the Two-sample KS test is used to check if there is a difference in these distribution in different periods. The test results are shown in Table 5.4.

Table 5.4: Two-sample KS test of distribution of proportion of passenger with TF 1, 2 and 3 in Period 1, 2 and 3 based on departure time

Period	*Statistic*	*First trip*						*Last trip*					
		Working day			*Non-working day*			*Working day*			*Non-working day*		
		TF 1	*TF 2*	*TF 3*	*TF 1*	*TF 2*	*TF 3*	*TF 1*	*TF 2*	*TF 3*	*TF 1*	*TF 2*	*TF 3*
Period 1 vs Period 2	KS statistic	0.16216	0.08108	0.16216	0.18919	0.05405	0.16216	0.16216	0.08108	0.10811	0.10811	0.10811	0.27027
	p-value	0.67590	0.99947	0.67590	0.47867	1.00000	0.67590	0.67590	0.99947	0.97495	0.97495	0.97495	0.11127
Period 2 vs Period 3	KS statistic	0.16216	0.13514	0.10811	0.10811	0.13514	0.05405	0.27027	0.21622	0.08108	0.08108	0.10811	0.13514
	p-value	0.67590	0.86307	0.97495	0.97495	0.86307	1.00000	0.11127	0.31362	0.99947	0.99947	0.97495	0.86307
Period 1 vs Period 3	KS statistic	0.08108	0.08108	0.13514	0.18919	0.13514	0.16216	0.16216	0.16216	0.10811	0.10811	0.13514	0.21622
	p-value	0.99947	0.99947	0.86307	0.47867	0.86307	0.67590	0.67590	0.67590	0.97495	0.97495	0.86307	0.31362

As can be seen from Table 5.4, all p-values are greater than 0.05, which means at a significance level of 0.05, there is no significant change in the departure time of first trip and last trip of passengers with TF 1, 2, and 3. This result is consistent with the observations of Figure 5.4, that is, no obvious differences occur.

5.4.3 Trip distance

In Figure 5.5, based on trip distance of each passenger, the distributions of proportion of passenger with TF 1/2/3 are highly coincident at different periods, and no change is observed. In order to verify this observation, the distributions of proportion of passengers of the same trip frequency at different periods were tested. The results are shown in Table 5.5.

Table 5.5: Two-sample KS test of distribution of proportion of passenger with TF 1, 2 and 3 in Period 1, 2 and 3 based on trip distance

Period	*Statistic*	*Working day*			*Non-working day*		
		TF 1	*TF 2*	*TF 3*	*TF 1*	*TF 2*	*TF 3*
Period 1 vs Period 2	KS statistic	0.06154	0.03077	0.07692	0.07692	0.04615	0.06154
	p-value	0.99950	1.00000	0.98765	0.98765	1.00000	0.99950
Period 2 vs Period 3	KS statistic	0.06154	0.06154	0.06154	0.04615	0.06154	0.04615
	p-value	0.99950	0.99950	0.99950	1.00000	0.99950	1.00000
Period 1 vs Period 3	KS statistic	0.04615	0.04615	0.07692	0.07692	0.06154	0.04615
	p-value	1.00000	1.00000	0.98765	0.98765	0.99950	1.00000

As can be seen from Table 5.5, all p-value values are greater than 0.05, that is, statistically speaking, there is no significant difference in the distribution of proportion of passengers with the same trip frequency at different periods. This is consistent with the observations.

5.4.4 Origin/destination area

Based on the selection of origin/destination area, Table 5.6 compares the percentage distribution of passengers with same trip frequency at different periods. From the results of the comparison, it can be seen that all p-values are greater than 0.05. Statistically, the percentage distribution of passengers of the same frequency does not change with time.

Table 5.6: Two-sample KS test of distribution of proportion of passenger with TF 1, 2 and 3 in Period 1, 2 and 3 based on origin/destination area

Period	*Statistic*	*Origin area*			*Destination area on working day*			*Destination area on non-working day*		
		TF 1	*TF 2*	*TF 3*	*TF 1*	*TF 2*	*TF 3*	*TF 1*	*TF 2*	*TF 3*
Period 1 vs Period 2	KS statistic	0.06780	0.05085	0.10169	0.05932	0.05932	0.11017	0.05932	0.05932	0.09322
	p-value	0.94158	0.99745	0.55264	0.98261	0.98261	0.44872	0.98261	0.98261	0.66331
Period 2 vs Period 3	KS statistic	0.05932	0.05932	0.05932	0.07627	0.06780	0.04237	0.04237	0.06780	0.05932
	p-value	0.98261	0.98261	0.98261	0.86941	0.94158	0.99990	0.99990	0.94158	0.98261
Period 1 vs Period 3	KS statistic	0.05932	0.05085	0.10169	0.06780	0.06780	0.11017	0.06780	0.07627	0.09322
	p-value	0.98261	0.99745	0.55264	0.94158	0.94158	0.44872	0.94158	0.86941	0.66331

Figure A.3, Figure A.4, Figure A.5 are passenger distributions for origin area and destination area on working/non-working day. Although statistical analysis shows that there is no significant change in the percentage distribution of passengers, from these three figures, it can be observed that there are distribution changes in the areas involved in individual stations.

From Figure A.3, for the origin area, the most obvious change is that by the time of Period 3, there were more passengers with TF 1 in the vicinity of Shenzhen North subway station. In Figure A.4, for the destination area on working day, the most obvious change was also found near the Shenzhen North subway station and was only seen in passengers with passengers with TF 1. Figure A.5 shows the passenger distribution on destination area on non-working day. From the distribution of passenger with TF 1 and 2, it can be observed that more passenger distributed in the vicinity of Shenzhen North subway station.

In order to understand the occurrence of distribution changes is the impact of accidental conditions, or the long-term impact of changes in station facilities or surrounding facilities, a qualitative analysis is conducted for this area to understand its changes in accessibility, attractiveness to passengers, and number of dwellings within the analysis timeframe.

Shenzhen North Subway Station is located in Longhua District, 9.3 km from downtown Shenzhen. It is the interchange station for Line 4 and Line 5. The station is an elevated station and is designed for integration with Shenzhen North High-Speed Rail (HSR) Station. It also has a long-distance bus terminal, a taxi departure station and a city bus departure station. Therefore, this is an important comprehensive transportation hub in the urban area of Shenzhen, named Shenzhen North Transportation Hub. In addition to providing transportation services, the hub's building is also equipped with commercial facilities.

First discuss the accessibility of the area. In this area, the Shenzhen North HSR Station was opened in December 2011. At the beginning of the opening, there were only trips to Guangzhou. From 2012 to 2014, the station continued to open new trains to other parts of China. The specific time is shown in Table 5.7:

Table 5.7: The start time and target location of HSR service in Shenzhen North Station (Yang, 2012)

Year	*Destination*
2012.4	Wuhan, Changsha
2012.9	Zhengzhou, Xi'an
2012.12	Beijing
2013.12	Xiamen
2014.9	Nanchang
2014.12	Guilin

The ever-increasing number of trains has brought a steady increase in passenger traffic to Shenzhen North. According to Chinese Media, by the end of 2014, the monthly passenger traffic of Shenzhen North HSR Station had reached 2 million.

The long-distance bus terminal was operated in April 2012. It mainly provides long-distance passenger services from Shenzhen North to surrounding provinces, Hong Kong and Macau. The designed passenger traffic is 6,000 passengers per day. After the operation of the high-speed rail station, the bus company began to set up docking stations in the hub center and opened new bus lines, mainly providing transportation services from Shenzhen to other administrative areas in the city.

Then discuss the attraction of the area to passengers. In the satellite image (Figure A.6), it can be observed that with the completion of the Shenzhen North Transportation Hub, there is a square in front of it. Confirmed by Google Maps, there are two commercial complexes in the square in front of the station, called Bingo Shopping Palaz and Youyue Time Square, which offer shopping, dining and leisure activities. They started to operate close to each other and opened some floors for trial operation around October 2014, so the impact on subway passengers should also be mainly experienced in the December 2014 data. Because during the observation period, only partial floors of the two commercial entities began to operate, and the operation time is short, the impact is limited.

As for the number of dwellings, it can be seen from the comparison of satellite maps in 2010 and 2012 that the buildings that were just built in 2012 are concentrated in Area 1 and Area 2, and these buildings are residential. Area 1 is a decorated low-cost housing with a total of 11,000 households, which was launched in January 2013 (Housing and Construction Bureau of Shenzhen Municipality, 2013). Area 2 has a total of 421 households and was launched in 2012.

In general, during the analysis timeframe, the area is improved in terms of accessibility, attractiveness to passengers, and the number of homes, which in theory can bring more travel needs. That is to say, the changes in the percentage of passengers distributed in the vicinity of the Shenzhen North Subway Station are not caused by sporadic factors, and these changes will persist for a long time. However, in the two-sample KS test, because the amount of data involved is large, and the areas with significant changes involve only a small amount of data, the test results show that there is no significant distribution difference for the entire data sample. And it should be noted here that the distribution of passengers involved in the Shenzhen North Subway Station cannot be ignored.

5.5 MEMBER SWITCHING ANALYSIS

Through statistical analysis, it can be found that passengers with the same trip frequency have similar PTCs in different periods. The reason for this may be that the trip frequency of most passengers is relatively stable, so the final reaction of PTCs has not changed. Another possible reason is that the passenger's trip frequency has changed, but their PTCs will also change accordingly, becoming the PTCs corresponding to the new trip frequency. For verification, Table 5.8 counts the trip frequency for each passenger at each period.

Table 5.8: Trip frequency of passengers in each period

Period 1	*Period 2*	*Period 3*	*# of passengers*	*Proportion*
TF1	TF1	TF1	225576	43.8%
TF1	TF1	TF2	83320	16.2%
TF1	TF1	TF3	12412	2.4%
TF1	TF2	TF1	79410	15.4%
TF1	TF2	TF2	72287	14.0%
TF1	TF2	TF3	13435	2.6%
TF1	TF3	TF1	7180	1.4%
TF1	TF3	TF2	10098	2.0%
TF1	TF3	TF3	11841	2.3%
Total			515559	100%

(a) TF 1 in Period 1 as reference

Period 1	*Period 2*	*Period 3*	*# of passenger*	*Proportion*
TF2	TF1	TF1	85893	22.0%
TF2	TF1	TF2	50181	12.9%
TF2	TF1	TF3	8053	2.1%
TF2	TF2	TF1	71784	18.4%
TF2	TF2	TF2	105156	27.0%
TF2	TF2	TF3	20800	5.3%
TF2	TF3	TF1	9512	2.4%
TF2	TF3	TF2	18114	4.6%
TF2	TF3	TF3	20562	5.3%
Total			390055	100.0%

(b) TF 2 in Period 1 as reference

Period 1	*Period 2*	*Period 3*	*# of passenger*	*Propotion*
TF3	TF1	TF1	10243	8.6%
TF3	TF1	TF2	6820	5.7%
TF3	TF1	TF3	2400	2.0%
TF3	TF2	TF1	11768	9.9%
TF3	TF2	TF2	18569	15.5%
TF3	TF2	TF3	8383	7.0%
TF3	TF3	TF1	8867	7.4%
TF3	TF3	TF2	17137	14.3%
TF3	TF3	TF3	35251	29.5%
Total			119438	100%

(c) TF 3 in Period 1 as reference

As can be seen from the results of Table 5.8, taking the passenger's trip frequency in Period 1 as a reference, for TF 1, 2 and 3, the proportion of passengers whose trip frequency remains the same is the highest. But the highest ratio is just over 40%. This shows that most of the passenger with TF 1/2/3 in Period 1 have changed their trip frequency in Period 3.

That is to say, the reason why the PTCs of TF 1, 2 and 3 showed no significant change in different periods is that, after the passengers change the trip frequency, their PTCs also change accordingly. For example, a passenger in Period 1 at TF 1, his daily ridership on non-working days is greater than the working day. When his trip frequency in Period 2 becomes TF 3, he would switch to have a larger daily ridership on working day.

5.6 SUMMARY

According to the results in this chapter, the following summary can be made:

- In the same period, the PTCs described by the four indicators show difference if the related passengers have different trip frequency.
- Statistically speaking, the PTCs related to passengers with TF 1/2/3 show no significant difference in Period 1, 2 and 3.
- Based on origin/destination area, the percentage distribution of passenger experienced changes near Shenzhen North Subway Station, and these changes involved varies among passengers with different trip frequencies.
- At different periods, a group of passengers with the same frequency of travel performed almost identical PTCs. But for this group of passengers the same trip frequency, most of the members experienced an update during each period

6 CONLUSION

6.1 FINDINGS

Based on SC data from Shenzhen Metro in December 2012, December 2013 and December 2014, this thesis describes passenger travel characteristics (PTCs) from temporal and spatial perspective by four indicators, daily ridership, departure time, trip distance and origin/destination area. According to the different trip frequency, the travel characteristics of the passengers with the trip frequencies of 1-5 (TF 1), 6-28 (TF 2) and 29-206 (TF 3) in the sample data in three different periods are respectively shown.

Here, the research question is:

What is the evolution of passenger characteristics based on SC data?

The research question is solved by answering the three sub-questions.

- **Based on SC data, how to describe the characteristics of passengers?**
  Considering the accuracy and completeness, four indicators that can directly obtained from SC data are selected from different perspectives to describe PTCs. They are daily ridership and departure time from temporal perspective, and trip distance and origin/destination area from spatial perspective.

  The daily ridership calculates the number of trips per passenger per day, which can understand that travel needs change over time. Departure time is the most common time for passengers to travel for the first trip and the last trip each day. It is used to understand the passengers' preferences and differences in travel time. Departure time on working days and non-working days are discussed separately.

  Trip distance calculates the travel distance most used by each passenger for all trips. This indicator reflects the mobility of passengers. It is also discussed here on working day and non-working days separately. The Origin/destination area is the most commonly used starting station for passengers' first trip per day and the most commonly used station for last trip. Different areas reveal the difference in passenger travel destinations and range of activities. The destination area is discussed on working and non-working days.

  The calculation of these four indicators avoids the impact of passengers' accidental behavior changes on the results.

- **How to identify the change of PTCs?**
  In order to find the change of PTCs, a longitudinal analysis is conducting, which is a method that allows the comparison of information obtained in the same indicator at different times from individual level. In the analysis process, combined with the research requirements, the requirements of the longitudinal analysis for the sample

data and the specific conditions of the Shenzhen Metro SC data, the data is first preprocessed. After the pre-processing is completed, the passengers involved in the sample data are characterized by the selected four indicators, and the PTCs of the sample passengers can be obtained. In order to better understand the difference between passengers, PTCs in each period are presented according to the passenger's trip frequency. Finally, the two-sample KS test is used to compare the quantitative data of PTCs in different periods, and it can be statistically understood whether there is a significant change. The evolution of PTCs can be understood by combining observations and statistical analysis.

- **What are the changes in PTCs during the analysis timeframe?**
  Based on the quantitative results involved in the four indicators, no change occurred in the PTCs associated with passengers with TF 1/2/3 at a significant level of 0.05. That is to say, statistically speaking, the PTCs involved in the sample selected in this study did not evolve during the analysis timeframe.

  However, according to the observation results, when the PTCs are discussed based on the origin/destination area, the percentage distribution of passengers is changed near the Shenzhen North Subway Station. The specific performance is as follows:

  – Origin area: More passengers with TF 1 were distributed near this station.

  – Destination area on working day: More passengers with TF 1 were distributed near this station.

  – Destination area on non-working day: More passengers with TF 1 and 2 were distributed near this station.

  In addition, although a group of passengers with the same trip frequency performed roughly the same PTCs at different period, most of the members in the group at different period were different.

It can be seen from the above results that the PTCs described by the four indicators have not evolved statistically from 2012 to 2014. However, when discussing based on the origin/destination area, the distribution of passengers at individual trip frequencies varies at individual stations, and these changes are related to the passenger's trip frequency. On this basis, the trip frequency of most passengers has changed, and the PTCs involved have also changed according to the new trip frequency. Eventually, as long as the travel frequency is the same, the PTCs presented at different period are roughly the same.

In conclusion, PTCs generated by the sample data in this thesis showed a slight evolution from 2012 to 2014, and passengers involved in the sample data with different travel frequencies show different PTCs evolution. This also proves that the trip frequency can be used to find the difference in passenger's PTCs evolution. In addition, not all indicators will show change during the evolution process.

6.2 SCIENTIFIC CONTRIBUTION

This thesis fills the scientific gap about exploring the PTCs evolution of a group of passengers while showing the difference of PTCs evolution among those passengers. In the previous studies, neither of these points was considered at the same time. Moreover, this study also demonstrates the possibility of exploring the difference of PTCs evolution through trip frequency.

6.3 LIMITATION

In this study, four indicators were selected to describe PTCs. In reality, the characteristics of passengers on travel are complex, and the evolution of PTCs cannot be fully described by these four indicators. Therefore, the evolution of PTCs discussed in this study can only be targeted at the four indicators selected.

In addition, the longitudinal analysis requires a consistent sample for the analysis timeframe, so the discussion of the evolution of the resulting PTCs is only for the passengers in the sample and does not represent the entire population. The sample data selected in this study only involved less than 10% of the smart cards, and most of the data was not used.

The results of this study show that the PTCs exhibited by the sample data show only a slight evolution trend. But at the same time, the speed of urban development in Shenzhen is relatively fast. From 2012 to 2014, the permanent registered population grew at an average rate of 7.5%, and also had a large permanent non-registered population (three times the resident population) which have a higher mobility (Shenzhen Statistics Bureau, 2018). In the absence of improvement in the subway network, in the face of urban development and higher mobility of residents, PTCs should have undergone a large degree of change, but the research results are different from expectations. There are two reasons for this. One is that the selected indicator is not suitable for describing PTCs evolution, and the other is that the sample of passengers involved in the sample data has relatively stable PTCs.

6.4 FURTHER ANALYSIS

Based on the limitations of this thesis, studies that can be conducted in the future are:

- Of the four indicators selected for the description of PTCs in this study, not all of them show changes during the evolution process. And when the passengers have different trip frequencies, the indicators which can show changes are different. Therefore, in the further research, it is interesting to explore whether the changes between these indicators are independent or interactive, and how the different trip frequencies of passengers are related to the changes of these indicators.
- 90% of the smart card datas in this thesis has not been covered. Because of the data preprocessing method, the users of this smart card have a higher changeability. For

example, they may have a high trip frequency in Period 1, but they no longer take the subway in Period 2. Therefore, this part of the passengers may show PTCs evolution with more obvious trends, and future research can be carried based on them.

BIBLIOGRAPHY

2013 annual statistical analysis report of urban rail transit in china [Computer software manual]. (2014). Retrieved from http://www.camet.org.cn/index.php?m=content&c=index&a=show&catid=42&id=3442

2014 annual statistical analysis report of urban rail transit in china [Computer software manual]. (2015). Retrieved from http://www.camet.org.cn/index.php?m=content&c=index&a=show&catid=18&id=1629

Agard, B., Morency, C., & Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, *39*(3), 399–404.

Bhaskar, A., Chung, E., et al. (2014). Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, *16*(3), 1537–1548.

Briand, A.-S., Côme, E., Trépanier, M., & Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, *79*, 274 - 289. Retrieved from http://www.sciencedirect.com/science/article/pii/S0968090X17301055 doi: https://doi.org/10.1016/j.trc.2017.03.021

Chu, K. K. A. (2015). Two-year worth of smart card transaction data–extracting longitudinal observations for the understanding of travel behaviour. *Transportation Research Procedia*, *11*, 365–380.

Deschaintres, E., Morency, C., & Trépanier, M. (2019). Analyzing transit user behavior with 51 weeks of smart card data. *Transportation Research Record*, 0361198119834917.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

Gong, Y., Lin, Y., & Duan, Z. (2017). Exploring the spatiotemporal structure of dynamic urban space using metro smart card records. *Computers, Environment and Urban Systems*, *64*, 169–183.

Goulet-Langlois, G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, *64*, 1–16.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., & González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, *151*(1-2), 304–318.

Hodges, J. (1958). The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, *3*(5), 469–486.

Housing and Construction Bureau of Shenzhen Municipality. (2013). *Longyueju and other low-income housing rental acceptance circular.* Retrieved from https://web.archive.org/web/20141022070503/http://www.szjs.gov.cn/ztfw/zfbz/fpcx/201301/t20130131_2105518.htm

Huang, J., Levinson, D., Wang, J., Zhou, J., & Wang, Z.-j. (2018). Tracking job and housing dynamics with smartcard data. *Proceedings of the National Academy of Sciences*, *115*(50), 12710–12715.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). *SciPy: Open source scientific tools for Python.* Retrieved from http://www.scipy.org/ ([Online; accessed ¡today¿])

Lee, S. G., & Hickman, M. (2013). Are transit trips symmetrical in time and space? evidence from the twin cities. *Transportation Research Record*, *2382*(1), 173–180.

Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, *58*, 135–145.

Morency, C., Trepanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, *14*(3), 193–203.

Ortega-Tong, M. A. (2013a). *Classification of london's public transport users using smart card data* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Ortega-Tong, M. A. (2013b). Classification of london's public transport users using smart card data..

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557–568.

Shenzhen Metro Group Co., Ltd. (n.d.). *Shenzhen metro introduction.* http://www.szmc.net/. (Accessed:2019-06-05)

Shenzhen municipal people's government. (n.d.). *Shenzhen city introduction.* http://www.sz.gov.cn/cn/. (Accessed:2019-06-05)

Shenzhen municipal transportation commission. (2016). *Shenzhen's comprehensive transportation "13th five-year plan".* http://jtys.sz.gov.cn/ydmh/xxgk/ghjh/fzgh/201702/P020170919652029168179.pdf. (Accessed:2019-06-05)

Shenzhen Statistics Bureau. (2018). *Shenzhen statistical yearbook.* Retrieved from http://www.sz.gov.cn/cn/xxgk/zfxxgj/tjsj/tjnj/201812/t20181229_14966437.htm

Shenzhen Tong Limited. (2013). *Shenzhen tong introduction.* https://www.shenzhentong.com/index.html. (Accessed:2019-06-05)

Southern Urban Daily. (2014). *Line 4 trains are lengthening.*

Tian Zongxing, L. G. (2018). Urban renewal strategy based on tod: A case study of longhua district, shenzhen. *Urban Planning International*, *33*(5), 93–98.

Tourangeau, R., Zimowski, M., & Ghadialy, R. (1997). *An introduction to panel surveys in transportation studies* (Tech. Rep.). United States. Federal Highway Administration.

Viallard, A., Trépanier, M., & Morency, C. (2019). Assessing the evolution of transit user behavior from smart card data. *Transportation Research Record*, 0361198119834561.

Wang, Z.-j., Chen, F., Wang, B., & Huang, J.-l. (2018, Sep 01). Passengers' response to transit fare change: an ex post appraisal using smart card data. *Transportation*, *45*(5), 1559–1578. Retrieved from https://doi.org/10.1007/s11116-017-9775-1 doi: 10.1007/s11116-017-9775-1

Yang, J. (2012). Practice and innovation of integrated construction of shenzhen north station. *China Construction*(7), 164–165.

Zhou, M., Wang, D., Li, Q., Yue, Y., Tu, W., & Cao, R. (2017). Impacts of weather on public transport ridership: Results from mining data from different sources. *Transportation research part C: emerging technologies*, *75*, 17–29.

A APPENDIX

A.1 ORIGIN/DESTINATION AREA

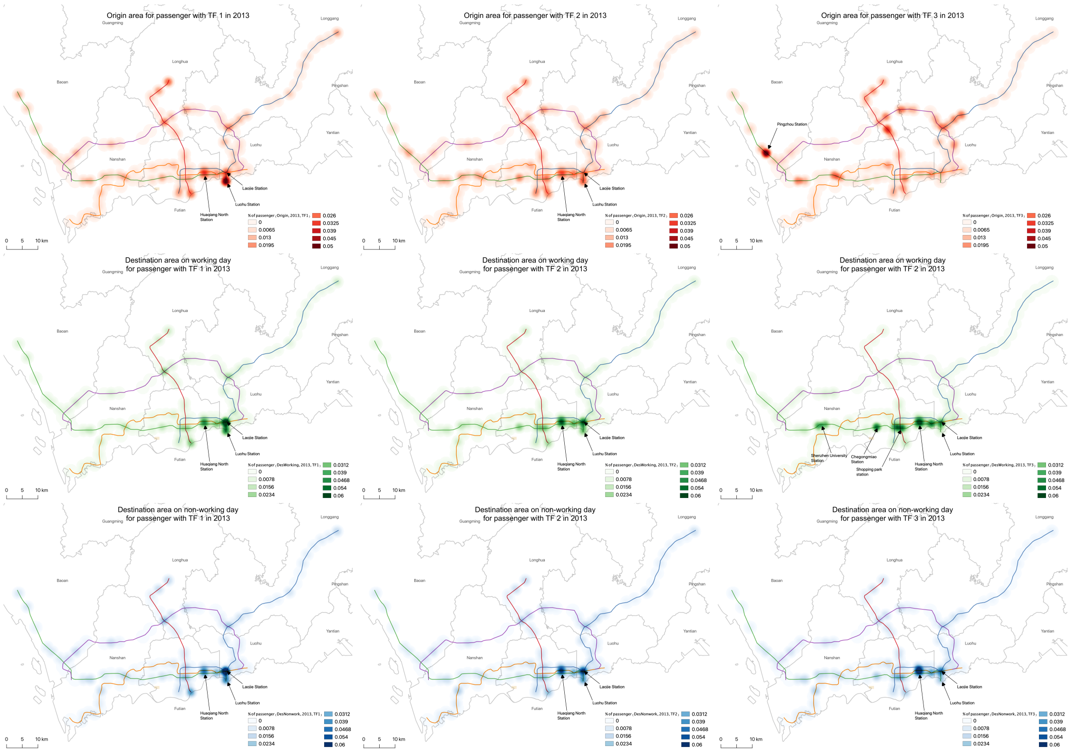

Origin area for passenger with TF 1 in 2013
Origin area for passenger with TF 2 in 2013
Origin area for passenger with TF 3 in 2013
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Pingzhou Station
Laojie Station
Luohu Station
Huaqiang North Station
% of passenger (Origin, 2013, TF1)
% of passenger (Origin, 2013, TF2)
% of passenger (Origin, 2013, TF3)
0
0.0065
0.013
0.0195
0.026
0.0325
0.039
0.045
0.05
0
5
10 km
Destination area on working day for passenger with TF 1 in 2013
Destination area on working day for passenger with TF 2 in 2013
Destination area on working day for passenger with TF 2 in 2013
Shenzhen University Station
Chegongmiao Station
Shopping park station
% of passenger (DesWorking, 2013, TF1)
% of passenger (DesWorking, 2013, TF2)
% of passenger (DesWorking, 2013, TF3)
0
0.0078
0.0156
0.0234
0.0312
0.039
0.0468
0.054
0.06
Destination area on non-working day for passenger with TF 1 in 2013
Destination area on non-working day for passenger with TF 2 in 2013
Destination area on non-working day for passenger with TF 3 in 2013
% of passenger (DesNonwork, 2013, TF1)
% of passenger (DesNonwork, 2013, TF2)
% of passenger (DesNonwork, 2013, TF3)

Figure A.1: The origin/destination area for passengers with different trip frequency in Period 2

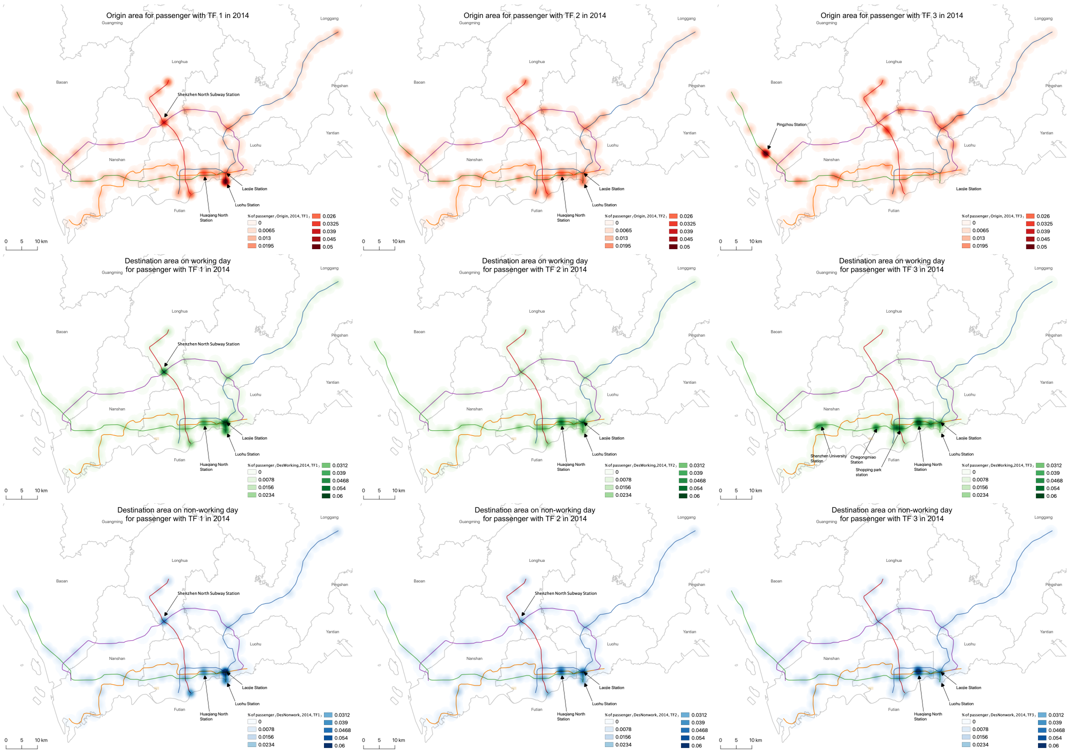

Origin area for passenger with TF 1 in 2014
Origin area for passenger with TF 2 in 2014
Origin area for passenger with TF 3 in 2014
Destination area on working day for passenger with TF 1 in 2014
Destination area on working day for passenger with TF 2 in 2014
Destination area on working day for passenger with TF 3 in 2014
Destination area on non-working day for passenger with TF 1 in 2014
Destination area on non-working day for passenger with TF 2 in 2014
Destination area on non-working day for passenger with TF 3 in 2014
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Shenzhen North Subway Station
Pingzhou Station
Laojie Station
Luohu Station
Huaqiang North Station
Shenzhen University Station
Chegongmiao Station
Shopping park station
% of passenger (Origin, 2014, TF1)
% of passenger (Origin, 2014, TF2)
% of passenger (Origin, 2014, TF3)
0
0.0065
0.013
0.0195
0.026
0.0325
0.039
0.045
0.05
% of passenger (DesWorking, 2014, TF1)
% of passenger (DesWorking, 2014, TF2)
% of passenger (DesWorking, 2014, TF3)
% of passenger (DesNonwork, 2014, TF1)
% of passenger (DesNonwork, 2014, TF2)
% of passenger (DesNonwork, 2014, TF3)
0
0.0078
0.0156
0.0234
0.0312
0.039
0.0468
0.054
0.06
0 5 10 km

Figure A.2: The origin/destination area for passengers with different trip frequency in Period 3

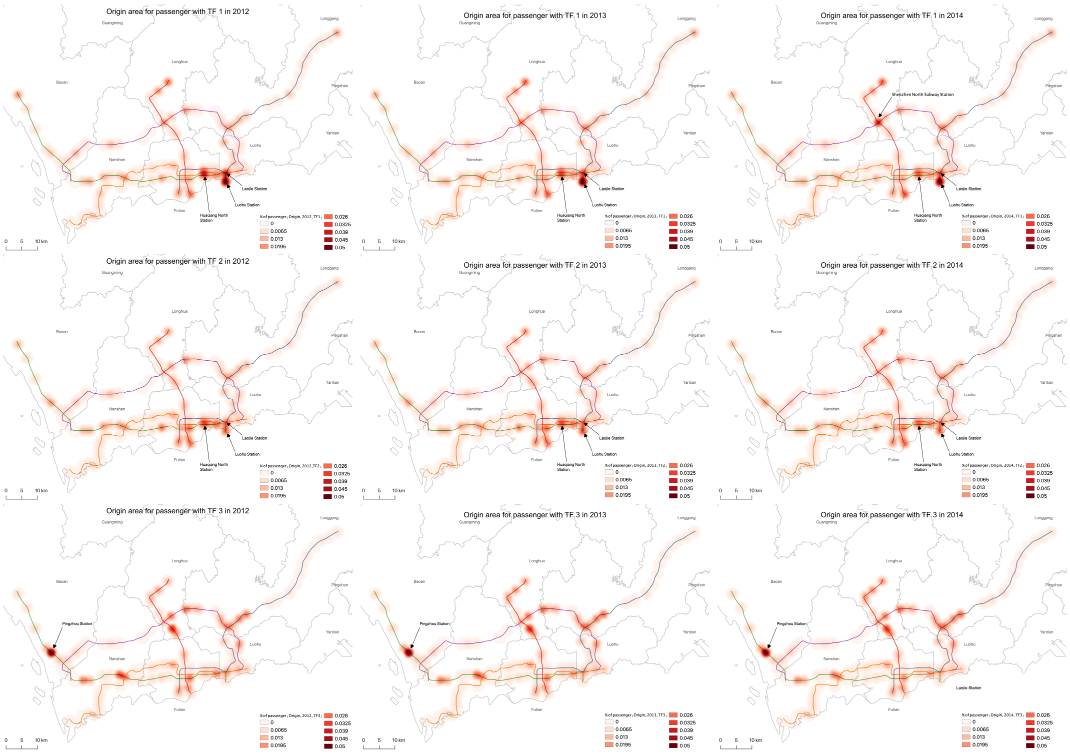

Origin area for passenger with TF 1 in 2012
Origin area for passenger with TF 1 in 2013
Origin area for passenger with TF 1 in 2014
Origin area for passenger with TF 2 in 2012
Origin area for passenger with TF 2 in 2013
Origin area for passenger with TF 2 in 2014
Origin area for passenger with TF 3 in 2012
Origin area for passenger with TF 3 in 2013
Origin area for passenger with TF 3 in 2014
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Shenzhen North Subway Station
Pingzhou Station
Laojie Station
Luohu Station
Huaqiang North Station
% of passenger (Origin, 2012, TF1)
% of passenger (Origin, 2013, TF1)
% of passenger (Origin, 2014, TF1)
% of passenger (Origin, 2012, TF2)
% of passenger (Origin, 2013, TF2)
% of passenger (Origin, 2014, TF2)
% of passenger (Origin, 2012, TF3)
% of passenger (Origin, 2013, TF3)
% of passenger (Origin, 2014, TF3)
0
0.0065
0.013
0.0195
0.026
0.0325
0.039
0.045
0.05
0 5 10 km

Figure A.3: The origin area for passengers with different trip frequency

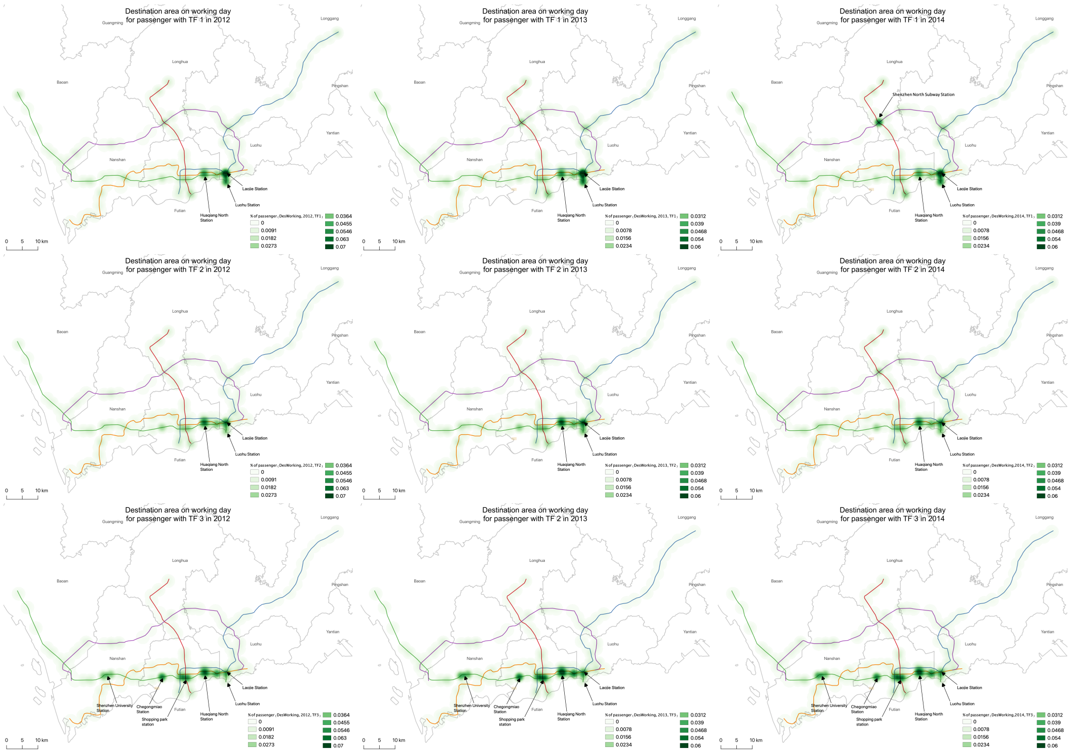

Destination area on working day for passenger with TF 1 in 2012
Destination area on working day for passenger with TF 1 in 2013
Destination area on working day for passenger with TF 1 in 2014
Destination area on working day for passenger with TF 2 in 2012
Destination area on working day for passenger with TF 2 in 2013
Destination area on working day for passenger with TF 2 in 2014
Destination area on working day for passenger with TF 3 in 2012
Destination area on working day for passenger with TF 2 in 2013
Destination area on working day for passenger with TF 3 in 2014
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Shenzhen North Subway Station
Laojie Station
Luohu Station
Huaqiang North Station
Shenzhen University Station
Chegongmiao Station
Shopping park station
% of passenger (DesWorking, 2012, TF1)
% of passenger (DesWorking, 2013, TF1)
% of passenger (DesWorking,2014, TF1)
% of passenger (DesWorking, 2012, TF2)
% of passenger (DesWorking, 2013, TF2)
% of passenger (DesWorking,2014, TF2)
% of passenger (DesWorking, 2012, TF3)
% of passenger (DesWorking, 2013, TF3)
% of passenger (DesWorking,2014, TF3)
0
0.0091
0.0182
0.0273
0.0364
0.0455
0.0546
0.063
0.07
0.0078
0.0156
0.0234
0.0312
0.039
0.0468
0.054
0.06
0 5 10 km

Figure A.4: The destination area on working day for passengers with different trip frequency

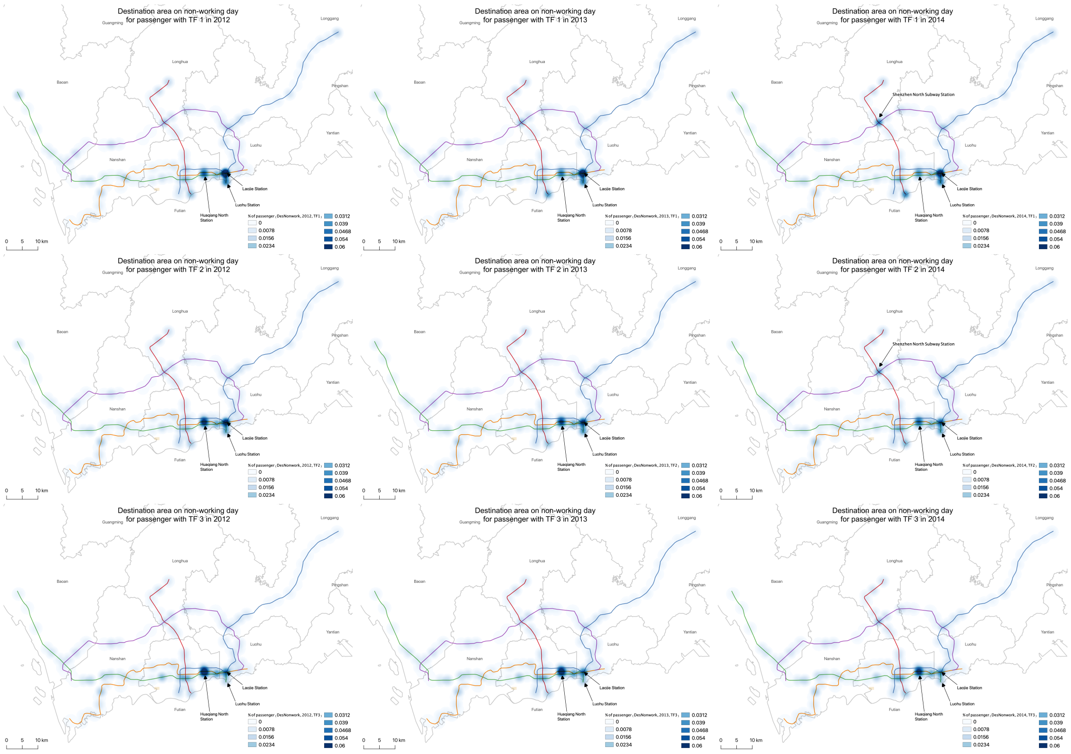

Destination area on non-working day for passenger with TF 1 in 2012
Destination area on non-working day for passenger with TF 1 in 2013
Destination area on non-working day for passenger with TF 1 in 2014
Destination area on non-working day for passenger with TF 2 in 2012
Destination area on non-working day for passenger with TF 2 in 2013
Destination area on non-working day for passenger with TF 2 in 2014
Destination area on non-working day for passenger with TF 3 in 2012
Destination area on non-working day for passenger with TF 3 in 2013
Destination area on non-working day for passenger with TF 3 in 2014
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Shenzhen North Subway Station
Laojie Station
Luohu Station
Huaqiang North Station
% of passenger (DesNonwork, 2012, TF1)
% of passenger (DesNonwork, 2013, TF1)
% of passenger (DesNonwork, 2014, TF1)
% of passenger (DesNonwork, 2012, TF2)
% of passenger (DesNonwork, 2013, TF2)
% of passenger (DesNonwork, 2014, TF2)
% of passenger (DesNonwork, 2012, TF3)
% of passenger (DesNonwork, 2013, TF3)
% of passenger (DesNonwork, 2014, TF3)
0
0.0078
0.0156
0.0234
0.0312
0.039
0.0468
0.054
0.06
0 5 10 km

Figure A.5: The destination area on non-working day for passengers with different trip frequency

A.2 SATELLITE IMAGE OF SHENZHEN NORTH SUBWAY STATION

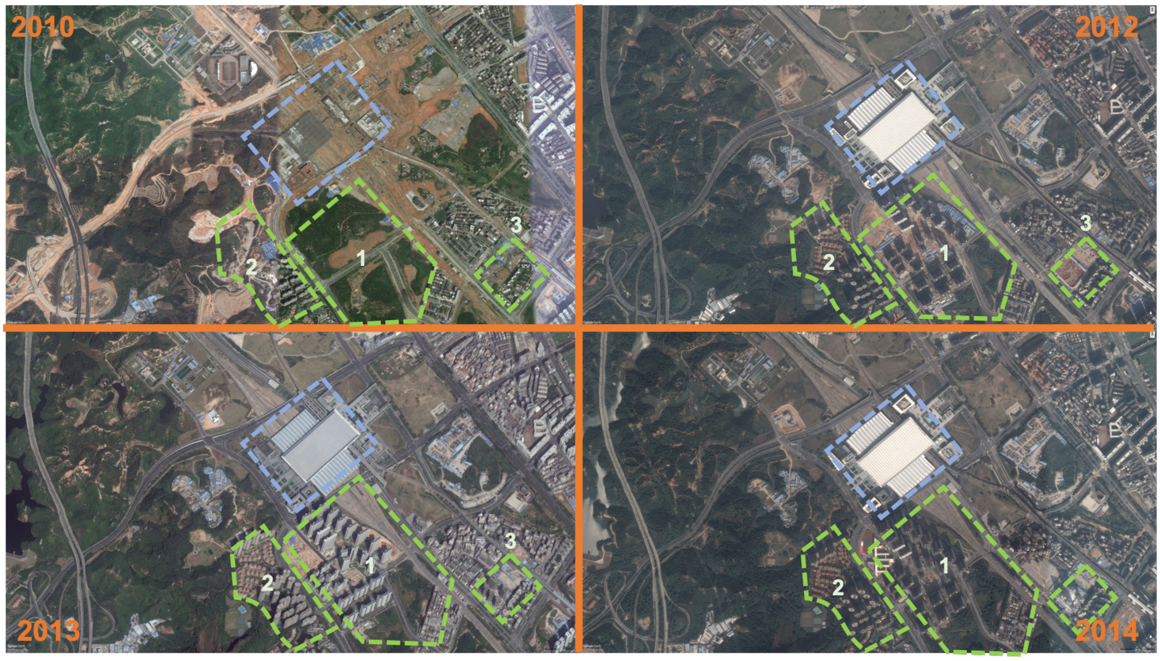


Figure A.6: Satellite image of shenzhen north subway station in 2010, 2012, 2013 and 2014

A.3 SCIENTIFIC PAPER

Exploring the Evolution of Passenger Characteristics Based on Smart Card Data: A Case Study of Shenzhen, China

Jijia Que,[*] Ding Luo, [†] Wei Pan, [‡] Rob van Nes, [§] Hans van Lint, [¶]

Delft University of Technology, Delft, The Netherlands

Abstract

Smart card data contains a large amount of passenger travel information, so it can be used to analyze passenger characteristics. Combined with smart card data from different periods, the evolution of passenger travel characteristics (PTCs) can be reflected. However, previous studies only analyzed the overall PTCs evolution shown by the sample data, or only analyzed the changes in PTCs of each passenger at the individual level without considering the overall situation. From temporal and spatial perspective, this study explored the PTCs evolution of a group of passengers, while classifying the passenger by trip frequency to show the difference of PTCs evolution among those passengers. The results show that passengers with different trip frequencies exhibit different evolution of PTCs.

Keywords: Smart card data, Passenger travel characteristics, Trip frequency, Evolution

1 Introduction

Passenger travel characteristics (PTCs) are characteristics that are sought from the information contained in trips, such as travel frequency, travel mode, departure location, and etc. Understanding PTCs can help develop passenger-oriented planning and service policies, so analysis of PTCs is a topic of constant interest to researchers and transportation service providers. However, PTCs vary with the change of many factors, such as network, price policy, service. The process or trend of PTCs change, which is called the evolution of PTCs, are useful for assessing the impact of external factors changes on passengers.

For a long time, the data used in related researches largely comes from traditional collection method, such as travel survey, which is time-consuming and expensive. With the development of technology, automatic toll collection devices are commonly used in transportation systems, such as automatic fare system (AFS). With AFS, the passenger should pay through the smart card (SC) each time they check-in and check-out In the process, a large quality of individual travel information can be recorded including card ID, boarding time, transaction type, boarding stations, etc., all the information generated after using SC is called SC data. SC data has a great potential to analysis passenger. Because each card has a unique ID, tracking ID can obtain the travel information generated by the card during a certain period of time, so that the travel characteristics of the passenger corresponding to the card can be known. The emergence of SC data avoids the data errors inolved in traditional collection methods and it is relatively inexpensive.

Recognizing the potential of SC data, some studies have demonstrated the feasibility of using SC data to explore PTCs evolution, for example [3, 8, 13, 14]. In these studies, the exploration of the evolution of PTCs is based on the transaction data within the analysis timeframe generated by a fixed passenger group, that is, a sample data is determined. There are two main methods to explore evolution. The first method first determines representative PTCs based on all transaction data in the sample. The analysis timeframe is then divided into consecutive periods that do not overlap, and the transaction data of the passengers in each period is selected. Based on the data for each period, explore each passenger's characteristics at different periods meets which representative PTCs. Finally, by tracking the representative PTCs of each passenger during each period, the PTCs evolution of the sample is known. The study using this method is [3], and the benefit is that the difference in evolution from different passengers can be known. Its limitation is that the representative PTCs

[*] MSc student, Department of Transport Intrastructure & Logistics, Faculty of Civil Engineering and Geosciences, Delft University of Technology

[†] Doctorial student, Department of Transport & Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology

[‡] Assistant Professor, Robot Dynamics Section, Department of Cognitive Robotics, Delft University of Technology

[§] Associate professor, Department of Transport & Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology

[¶] Professor, Department of Transport & Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology calculated here are based on all sample data, that is, they are unchanged during the analysis timeframe, and the evolution of PTCs exhibited by all sample passengers cannot be observed. The second method is to first divide the analysis timeframe into consecutive periods that do not overlap, and find the PTCs in each period. Learn about evolution by comparing PTCs from different periods. The PTCs for each period obtained by this method are all characterized by the appearance of all passengers involved in each period. Studies in this method are [4, 8, 13, 14]. Such an method allows researchers to understand the different PTCs of a fixed group of passengers at different periods, but cannot explore the differences in evolution between different passengers within the group.

In general, it is feasible to explore the evolution of PTCs through SC data, but no relevant research considers the changes in PTCs exhibited by all sample passengers, and also considers the difference in PTCs evolution between different passengers. Understanding the evolution of PTCs across the entire passenger group and the evolution of different passenger PTCs in the group can help to know the impact of changes in policies or services on passengers from different levels. This thesis aims at exploring the evolution of PTCs and the difference of the evolution among passengers. The analysis is accomplished based on the SC card data from Shenzhen Metro, which provides complete historical records of all passengers' transactions in December 2012, December 2013 and December 2014.

The rest part of this thesis is structured as follows. Chapter 2 is literature review. Chapter 3 is methodology, which introduces how to describe PTCs from SC data and how to find the change of PTCs. Chapter 4 is the research background and data description. Chapter 5 shows the results of PTCs for passengers in different periods and discuss the PTCs change. Chapter 6 summarizes the main findings.

2 Literature review

This section would conduct literature review to understand how passenger characterization and classification were carried out in previous studies. In addition, the literature focusing on PTCs evolution would be summarized to understand the method they have taken in exploring evolution.

2.1 Passenger characterization based on SC data

From the existing research on analyzing passengers through SC data, the researcher's description of PTCs is from different perspectives according to the research requirements. From these perspectives, different indicators can be used for quantification. There are four perspectives are summarized here that can be used to describe the characteristics of passengers, which is temporal perspective, spatial perspective, activity perspective and other perspective.

Transaction time is an important information contained in the SC data, so investigating this dynamic from temporal perspective become a possible choice. In the study of [10], the author explores the temporal characteristics by number of traveling day and number of similar departure times for a week. In a study based on SC data from Beijing metro, travel frequency is used to classify commuter and non-commuter [8], which is number of trips performed per week. From temporal perspective, it also can be discussed in a shorter time scale, like within one day. The time is divided according to the prescribed interval, for example, 24 hours a day, divided into 1-hour interval [2], and then the temporal characteristics identified by the time selection. In addition to the above indicators, time-related indicators include ridership [4] and trip number [1].

Since the transactions in SC data include trading locations, it is possible to explore passengers' spatial characteristics. One example is the regularity of station, which is used to find the home and job station according to the most used station [8]. After knowing the home/job station of each passenger, the passenger distribution on home/job station can be visualized combined with the geographic information of each station. However, the drawback of this method has been recognized [9]. In addition to using trading location directly, some indicators calculated in combination with other databases can also be used to describe spatial characteristics, such as selected route and trip distance [6, 10].

In some studies, activity is another perspective used to describe PTCs. Activity duration is an example. It refers to the length of time a passenger has activities at a destination. Different lengths of time reflect different types of activities. For example, rest at home tends to take longer than working outside [11]. Combined with activity duration and some other information, such as home/job station, researchers infer different activity status of passengers. With knowing the type and duration of activities, it is possible to capture the activity sequence within which each trips occurs, some studies chose to describe and distinguish PTCs based on activity sequence [5].

In addition to the three perspectives mentioned above, there many related researches attempt to interpret PTCs from richer perspectives, such as passenger's social attributes [1, 3].

In order to make full use of the potential of SC data and a more complete understanding of PTCs, research is often described from multiple perspectives and using several indicators.

2.2 Passenger classification

The passenger classification is helpful to have a better understand of passengers [2] by analysing the typical characteristics of each group of passengers with similar travel behavior. Passengers classification can be based on personal characteristics, which can be indirectly found from card type [3]. Another way is to classify according to different travel characteristics. For example, [5] divides passengers with similar activity sequences into groups. In terms of accuracy, the information that can be obtained directly from SC data is more accurate, but the type of data is not much. For characteristics inferred from SC data, although its accuracy is yet to be verified, it provides a richer perspective on difference between passengers.

2.3 Evolution of PTCs

Through Google scholar, the following would introduce all relevant research retrieved from 2015 to the present. In the study of [3], ten representative time activities characteristics were obtained from the whole sample data. Then, based on the data of each year, passengers are grouped according to their closest typical time activities characteristics and the PTCs evolution by tracking the group each passenger belong to in each year. In the study of [8], the authors classify the passengers according to the changes in the home stations and work stations of these commuters within seven years. Afterwards, the authors analyzed the changes in average travel time and housing expenditure for each type of passenger. The study of [13], the authors classify passengers according to the data in first week and obtain representative initial characteristics. Then, based on whether there is a policy holiday and school break next week, adjust initial representative characteristics to the new representative characteristics. In this study, the authors study evolution through changes in representative characteristics each week. In the study of [4], the PTCs are discribed based on the data at each year/week/day generated by all cards in selected sample.

It is found that the exploration of PTCs evolution in existing research can be divided into two types. One is to group passengers according to the characteristics that are found from all sample data, and then study the belonging group of passengers in different periods, evolution here is represented by member switching (Figure 1); the other is to compare the overall characteristics of the target passenger group in different periods, that is, aggregate characteristics changes (Figure 2).

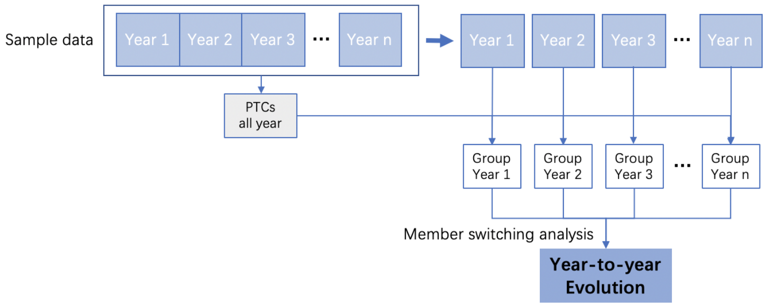


Figure 1: Evolution exploration method: member switching analysis (year-to-year for example)

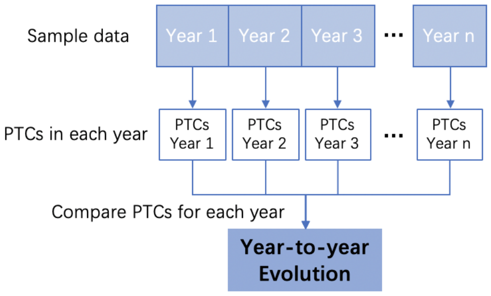


Figure 2: Evolution exploration method: aggregate characteristics change (year-to-year for example)

And a scientific gap about exploring the PTCs evolution of a group of passengers while showing the difference of PTCs evolution among thoes passengers is found here.

3 Methodology

Figure 3 shows the entire analysis process of this paper, which is mainly divided into four steps and are represented by squares in the figure. The raw data cannot be directly used in the analysis, so the data is first preprocessed by step one. Then, in the second step, all passengers are grouped in order to better observe the difference between the them. After the two steps, the research samples in different periods can be acquired. Then, from the temporal and spatial perspective, the selected indicators quantify the PTCs. After obtaining quantitative results of PTCs at different times, the evolution of the PTCs will be found by longitudinal analysis. The four steps are described in detail below.

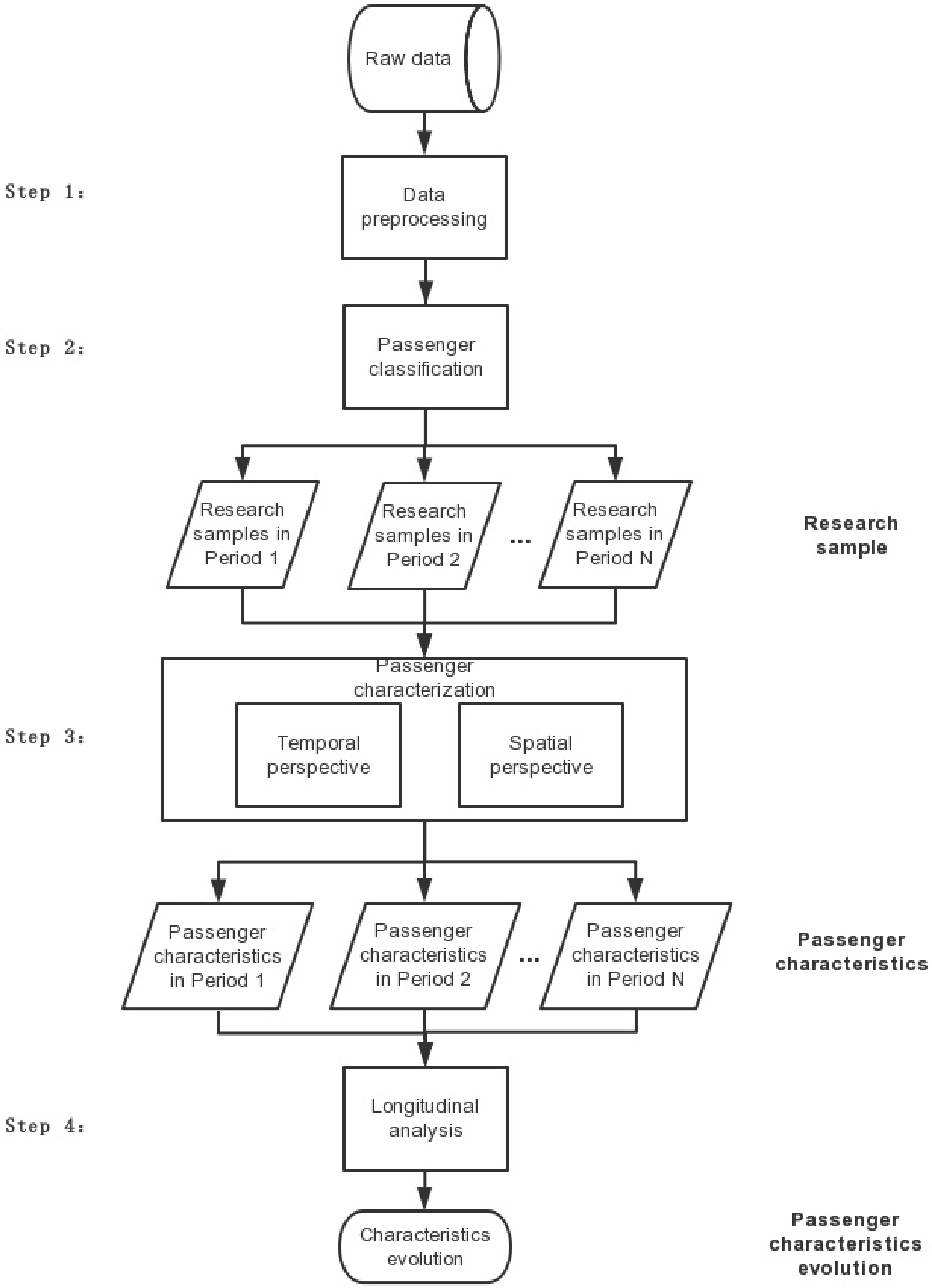

Raw data
Step 1:
Data preprocessing
Step 2:
Passenger classification
Research samples in Period 1
Research samples in Period 2
...
Research samples in Period N
Research sample
Step 3:
Passenger characterization
Temporal perspective
Spatial perspective
Passenger characteristics in Period 1
Passenger characteristics in Period 2
...
Passenger characteristics in Period N
Passenger characteristics
Step 4:
Longitudinal analysis
Characteristics evolution
Passenger characteristics evolution

Figure 3: Analysis process

In data preprocessing, incomplete data and erroneous data would be deleted first. Then, all relevant transactions will be sorted out by the ID of each card, which is the transactions of the passenger corresponding to this card. Only the cards with transaction records in all periods during the analysis timeframe can be retained.

Considering the accuracy and difficulty in obtaining data, the passenger classification indicator used in this paper is trip frequency, which can be obtained by calculating the number of transactions involved. And studies have shown that passengers with different trip frequencies have different characteristics [8]. Therefore, by grouping passengers according to trip frequency, it is more convenient to look for different PTCs evolutions.

After the passenger classification, the PTCs of each group in each period of analysis timeframe would be described separately by indicators from temporal and spatial perspective. The reason for choosing these two perspectives is that time- and space-related indicators can often be obtained directly from the SC data, and multi-perspective descriptions provide a more complete picture of PTCs. There two indicators related to temporal perspective are selected here, which are daily ridership and departure time. Trip distance and origin/destination area, is selected to describe PTCs from spatial perspective. The quantified results are: daily ridership, passenger proportion in each time interval based departure time (for first and last trip separately, for working day and non-working day separately), passenger proportion in each distance interval trip distance (for working day and non-working day separately) and passenger proportion near each station based on origin/destination area (only destination locations discuss on working day and non-working day separately).

Finally, for understanding the changes in characteristics, an longitudinal analysis in conducted. The specific method adopted is observation and statistical analysis. All the quantified results are visualized first in order to observe the difference directly. Observed results are subjective judgments from the investigator and there may be errors. The two-sample Kolmogorov–Smirnov test (KS test) will be used to analyze the observations, which is a nonparametric test and can be used to test whether two data samples come from the same continuous distribution [7].

4 Research background and data introduction

4.1 Research background

Shenzhen is first economic zone in China, which is an immigrant city and experienced explosive population growth. In order to meet the development needs of the city and the increasing travel demand of residents, Shenzhen's public transportation system has also experienced rapid development. The Shenzhen Metro was first opened in 2004. From 2012 to 2014, the network of Shenzhen Metro consisted by 5 operational metro lines with 118 stations, without any new lines and stations.

Depend on the issuer, tickets used in Shenzhen Metro can be divided into three types, that is identification, Shenzhen Tong and RFTD tokens [Shenzhen Metro Group Co., Ltd.]. And only the data generated by Shenzhen Tong can be used in longitudinal analysis. In order to facilitate the calculation, a unified operating time of Shenzhen Metro will be used in this study, that is 06:00-24:00.

4.2 SC data and network data

The raw data includes the relevant records of all passengers using Shenzhen Metro in December 2012, December 2013 and December 2014 for a total of three months. A record is generated each time a passenger uses Shenzhen Pass or Token to enter or exit the station. The types of data contained in each record are listed in Table 1.

Table 1: Data type

Card ID	***Card type***[1]	***Transaction tyepe***	***Station***	***Gate machine***[1]	***Transaction time***
987727139	98	21	1268012000	268012203	"2013-11-30,15:52:05"

[1] There is no card type and gate machine in the December 2012 record

The network data contains the name, ID and latitude and longitude of each station, as well as the number, geographical location and length of each line.

5 Result

5.1 Sample data

The sample data is determined according to the method in data preprocessing and the SC data from Shenzhen Metro. In the sample data, only records for four consecutive weeks in each month are retained, in order to ensure the comparability of the data in time for longitudinal analysis. Therefore, the analysis timeframe consists of three periods, and the daily operation time is 06:00-24:00 during each period. The three periods are:

- Period 1: December 3, 2012 to December 29, 2012

- Period 2: December 2, 2013 to December 28, 2013
- Period 3: December 1, 2014 to December 27, 2014

After deleting the records outside the analysis timeframe and the incomplete and erroneous records, only Shenzhen Tong, which has transaction records in all three periods, is retained. These Shenzhen Tong transaction records constitute the sample data used in this thesis. The number of Shenzhen Tong card involved in the sample data is 1,025,052, accounting for 7.2% of all cards, including the common card, student card and courtesy card. The distribution of the three different types of Shenzhen Tong is shown in Figure 4.

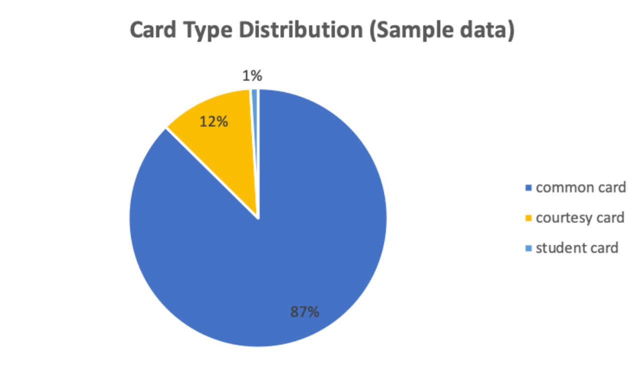


Figure 4: The card type distribution in sample data

5.2 Passenger classification

Figure 5 shows the distribution of passenger trip frequencies over different periods. It can be seen from the figure that the distribution of trip frequency is similar in the three periods, and the number of passengers in each interval within three periods is not much different. In sample data, about 500,000 passengers travel at a frequency of 1-5 times a month, accounting for half of the total number of people. As the trip frequency increases, the number of passengers in the corresponding interval decreases exponentially.

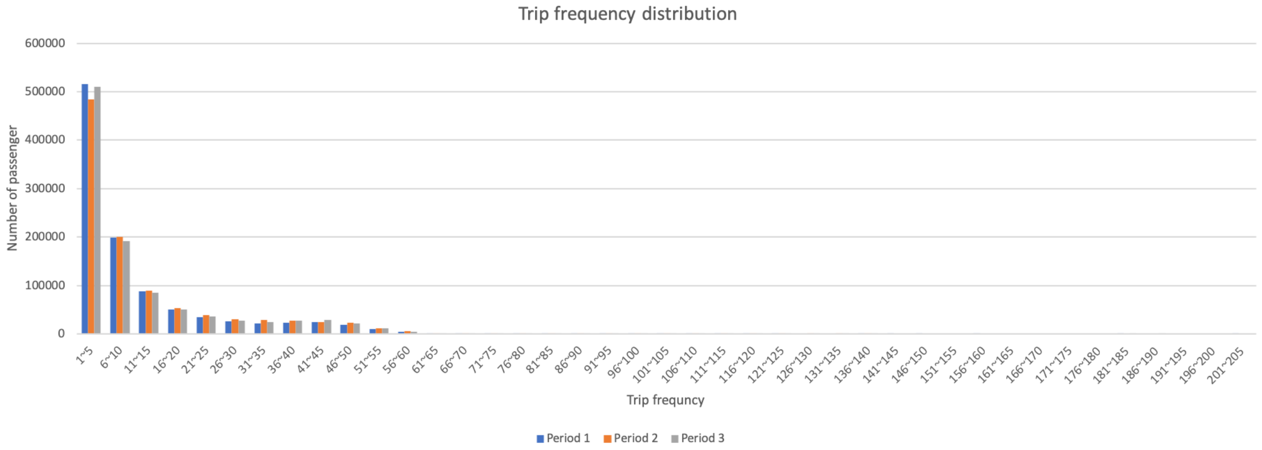


Figure 5: Trip frequency distribution during three periods in analysis timeframe

In order to catch the difference among all passengers as much as possible, considering the characteristic of passenger's trip frequency distribution mentioned above, the trip frequency in Figure 5 is divided into three intervals for the passenger classification. The details about the three interval and why they so divided are as follows:

- **Trip frequency in 1-5 (TF 1):** Although the frequency of their use of the subway is low, but the number of passengers is large, the total number of trips eventually generated is also considerable. Therefore, the characteristics of this part of the passengers deserve to be studied separately.
- **Trip frequency in 6-28 (TF 2):** Figure 5 shows that half of the passengers in the sample data are distributed in the range of 6-205, which is quite wide. It is reasonable to doubt that these passengers

should have significant different PTCs since their trip frequency has a greatly difference. Therefore, the travel frequency is divided into two intervals of 6-28 and 29-205. Among them, passengers with a travel frequency in the range of 6-28 have the average number of trips per day (28 days in each period in analysis timeframe) is less than or equal to one; Passengers in the interval 29-205 have the average number of trips per day is greater than one.

- **Trip frequency in 29-205 (TF 3):** Same as TF 2.

5.3 PTCs and difference among passengers

5.3.1 Daily ridership

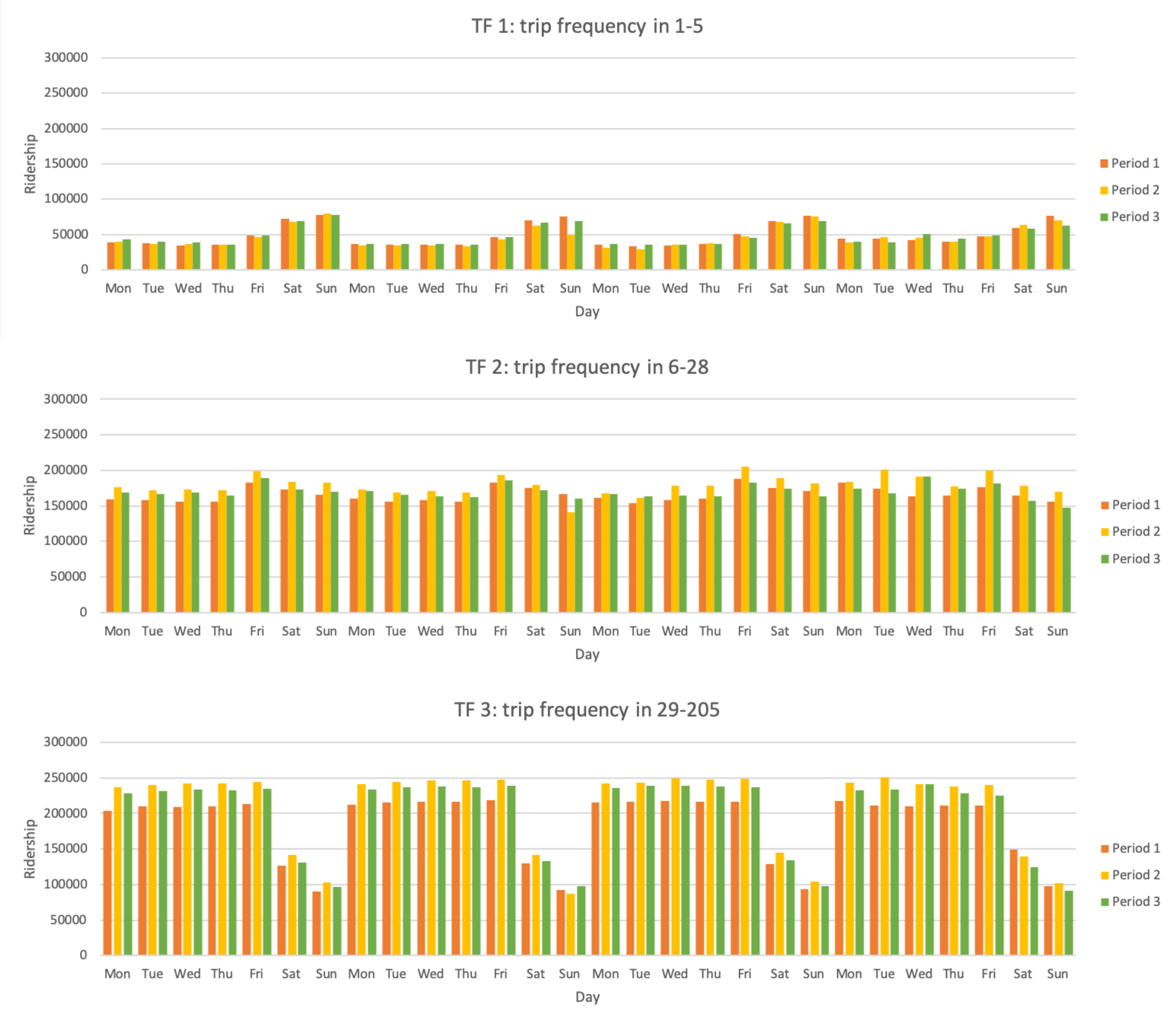


Figure 6: Daily ridership for passengers with different trip frequency in the three periods

Firstly, the daily ridership generated by passengers with three different trip frequencies in three different periods are summed separately, and then the proportion of daily ridership of passengers with different trip frequency is calculated separately. This is because the number of passengers with the same trip frequency in different periods is different, so the daily ridership of a certain period will be relatively more or less overall, and KS test may mistakenly think that the daily ridership distribution of these two periods is different. But the actual situation may only be caused by an increase in the number of people, and the passenger's travel needs have not changed. The percentage can be used to avoid such errors. Here, the proportion of daily ridership from passengers in each trip frequency is compared. The test results are shown in Table 2: As can be seen from Table 2, all p-values are greater than 0.05, which means that the null hypothesis cannot be rejected. This also shows that at a significance level of 0.05, which means the daily ridership proportion by passengers with TF 1/2/3 in the three periods obey the same distribution, and no significant changes occur.

Table 2: Two-sample KS test of proportion of daily ridership by passenger with TF 1, 2 and 3 in Period 1, 2 and 3

Period	***Trip frequency***	***KS statistic***	***p-value***
Period 1 vs Period 2	TF 1	0.10714	0.99503
	TF 2	0.14286	0.91681
	TF 3	0.25000	0.30035
Period 2 vs Period 3	TF 1	0.21429	0.49026
	TF 2	0.14286	0.91681
	TF 3	0.14286	0.91681
Period 1 vs Period 3	TF 1	0.28571	0.16875
	TF 2	0.25000	0.30035
	TF 3	0.32143	0.08756

However, as can be seen from Figure 6, the daily ridership of Period 2 is generally higher than the other two periods except for the second Sunday (December 15, 2013). In the real world, there are many situations in which abnormal ridership changes, including the occurrence of extreme weather, temporary traffic control, and large-scale gatherings. The sporadic effects of these external factors may cause changes in ridership of one day, but the ridership will also recover after the external conditions return to normal. Therefore, when comparing the distribution of different periods, such an irregular ridership changes of one day should be ignored. In other words, there is no significant change in the proportion of daily ridership of passengers with different travel frequencies.

5.3.2 Departure time

Based on each passenger's departure time of first and last trip, Figure 7 shows the distribution of the percentage of passengers with TF 1, 2 and 3 over on day. Here, the Two-sample KS test is used to check if there is a difference in these distribution in different periods. The test results are shown in Table 3.

Table 3: Two-sample KS test of distribution of proportion of passenger with TF 1, 2 and 3 in Period 1, 2 and 3 based on departure time

Period	***Statistic***	***First trip***						***Last trip***					
		Working day			***Non-working day***			***Working day***			***Non-working day***		
		TF 1	***TF 2***	***TF 3***	***TF 1***	***TF 2***	***TF 3***	***TF 1***	***TF 2***	***TF 3***	***TF 1***	***TF 2***	***TF 3***
Period 1 vs Period 2	KS statistic	0.16216	0.08108	0.16216	0.18919	0.05405	0.16216	0.16216	0.08108	0.10811	0.10811	0.10811	0.27027
	p-value	0.67590	0.99947	0.67590	0.47867	1.00000	0.67590	0.67590	0.99947	0.97495	0.97495	0.97495	0.11127
Period 2 vs Period 3	KS statistic	0.16216	0.13514	0.10811	0.10811	0.13514	0.05405	0.27027	0.21622	0.08108	0.08108	0.10811	0.13514
	p-value	0.67590	0.86307	0.97495	0.97495	0.86307	1.00000	0.11127	0.31362	0.99947	0.99947	0.97495	0.86307
Period 1 vs Period 3	KS statistic	0.08108	0.08108	0.13514	0.18919	0.13514	0.16216	0.16216	0.16216	0.10811	0.10811	0.13514	0.21622
	p-value	0.99947	0.99947	0.86307	0.47867	0.86307	0.67590	0.67590	0.67590	0.97495	0.97495	0.86307	0.31362

As can be seen from Table 3, all p-values are greater than 0.05, which means at a significance level of 0.05, there is no significant change in the departure time of first trip and last trip of passengers with TF 1, 2, and 3. This result is consistent with the observations of Figure 7, that is, no obvious differences occur.

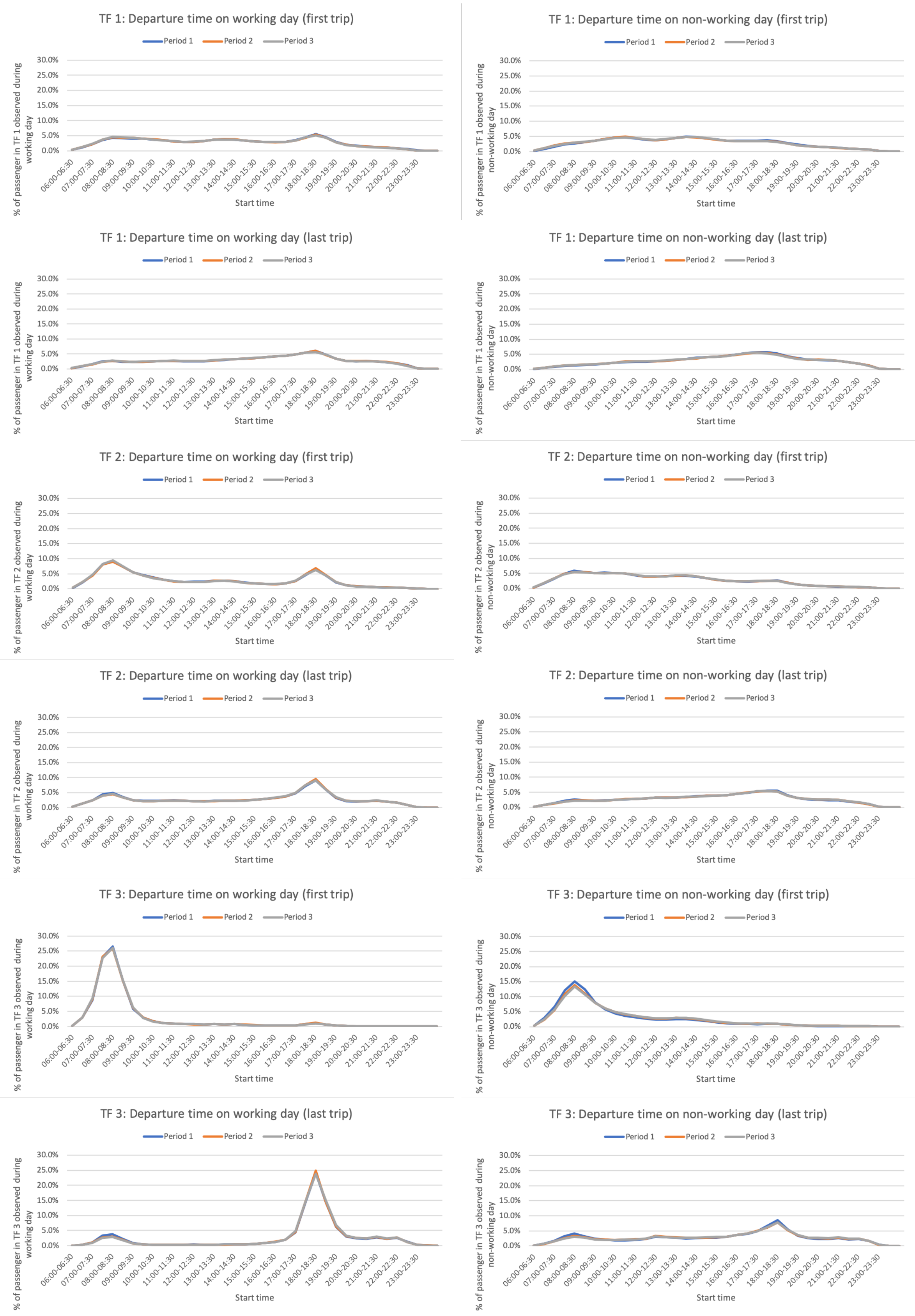


Figure 7: Departure time for passengers with different trip frequency in the three periods

5.3.3 Trip distance

Table 4: Two-sample KS test of distribution of proportion of passenger with TF 1, 2 and 3 in Period 1, 2 and 3 based on trip distance

Period	***Statistic***	***Working day***			***Non-working day***		
		TF 1	***TF 2***	***TF 3***	***TF 1***	***TF 2***	***TF 3***
Period 1 vs Period 2	KS statistic	0.06154	0.03077	0.07692	0.07692	0.04615	0.06154
	p-value	0.99950	1.00000	0.98765	0.98765	1.00000	0.99950
Period 2 vs Period 3	KS statistic	0.06154	0.06154	0.06154	0.04615	0.06154	0.04615
	p-value	0.99950	0.99950	0.99950	1.00000	0.99950	1.00000
Period 1 vs Period 3	KS statistic	0.04615	0.04615	0.07692	0.07692	0.06154	0.04615
	p-value	1.00000	1.00000	0.98765	0.98765	0.99950	1.00000

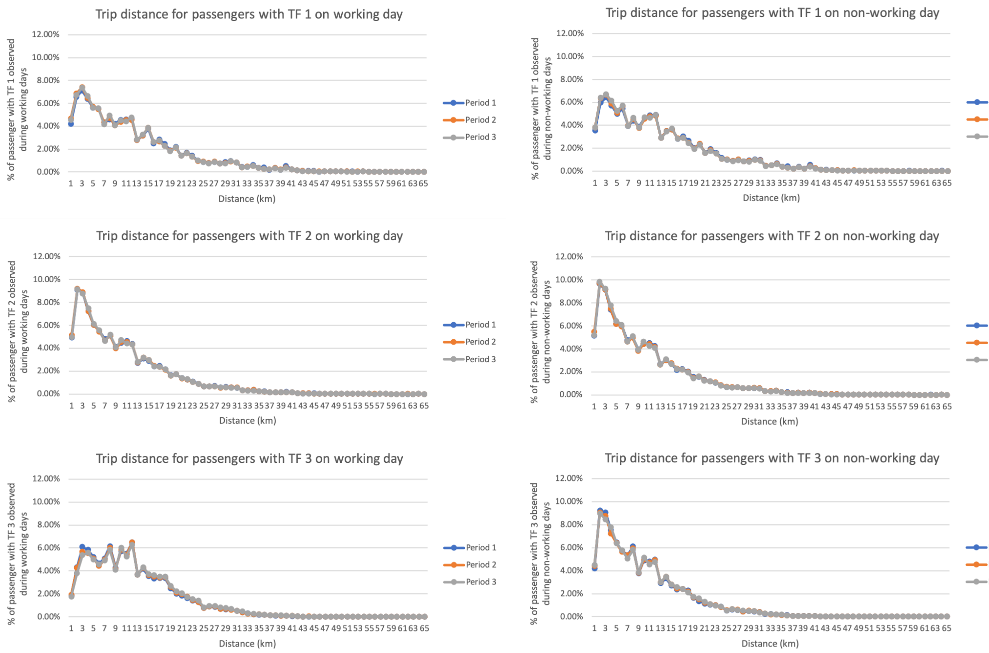


Figure 8: Trip distance for passengers with different trip frequency in the three periods

In Figure 8, based on trip distance of each passenger, the distributions of proportion of passenger with TF 1/2/3 are highly coincident at different periods, and no change is observed. In order to verify this observation, the distributions of proportion of passengers of the same trip frequency at different periods were tested. The results are shown in Table 4.

As can be seen from Table 4, all p-value values are greater than 0.05, that is, statistically speaking, there is no significant difference in the distribution of proportion of passengers with the same trip frequency at different periods. This is consistent with the observations.

5.3.4 Origin/destination area

Based on the selection of origin/destination area, Table 5 compares the percentage distribution of passengers with same trip frequency at different periods. From the results of the comparison, it can be seen that all p-values are greater than 0.05. Statistically, the percentage distribution of passengers of the same frequency does not change with time.

Figure 9, Figure 10, Figure 11 are passenger distributions for origin area and destination area on working/non-working day. Although statistical analysis shows that there is no significant change in the percentage distribution of passengers, from these three figures, it can be observed that there are distribution changes in the areas involved in individual stations.

From Figure 9, for the origin area, the most obvious change is that by the time of Period 3, there were more passengers with TF 1 in the vicinity of Shenzhen North subway station. In Figure 10, for the destination area on working day, the most obvious change was also found near the Shenzhen North subway station and was only seen in passengers with passengers with TF 1. Figure 11 shows the passenger distribution on destination

Table 5: Two-sample KS test of distribution of proportion of passenger with TF 1, 2 and 3 in Period 1, 2 and 3 based on origin/destination area

Period	***Statistic***	***Origin area***			***Destination area on working day***			***Destination area on non-working day***		
		TF 1	***TF 2***	***TF 3***	***TF 1***	***TF 2***	***TF 3***	***TF 1***	***TF 2***	***TF 3***
Period 1 vs Period 2	KS statistic	0.06780	0.05085	0.10169	0.05932	0.05932	0.11017	0.05932	0.05932	0.09322
	p-value	0.94158	0.99745	0.55264	0.98261	0.98261	0.44872	0.98261	0.98261	0.66331
Period 2 vs Period 3	KS statistic	0.05932	0.05932	0.05932	0.07627	0.06780	0.04237	0.04237	0.06780	0.05932
	p-value	0.98261	0.98261	0.98261	0.86941	0.94158	0.99990	0.99990	0.94158	0.98261
Period 1 vs Period 3	KS statistic	0.05932	0.05085	0.10169	0.06780	0.06780	0.11017	0.06780	0.07627	0.09322
	p-value	0.98261	0.99745	0.55264	0.94158	0.94158	0.44872	0.94158	0.86941	0.66331

area on non-working day. From the distribution of passenger with TF 1 and 2, it can be observed that more passenger distributed in the vicinity of Shenzhen North subway station.

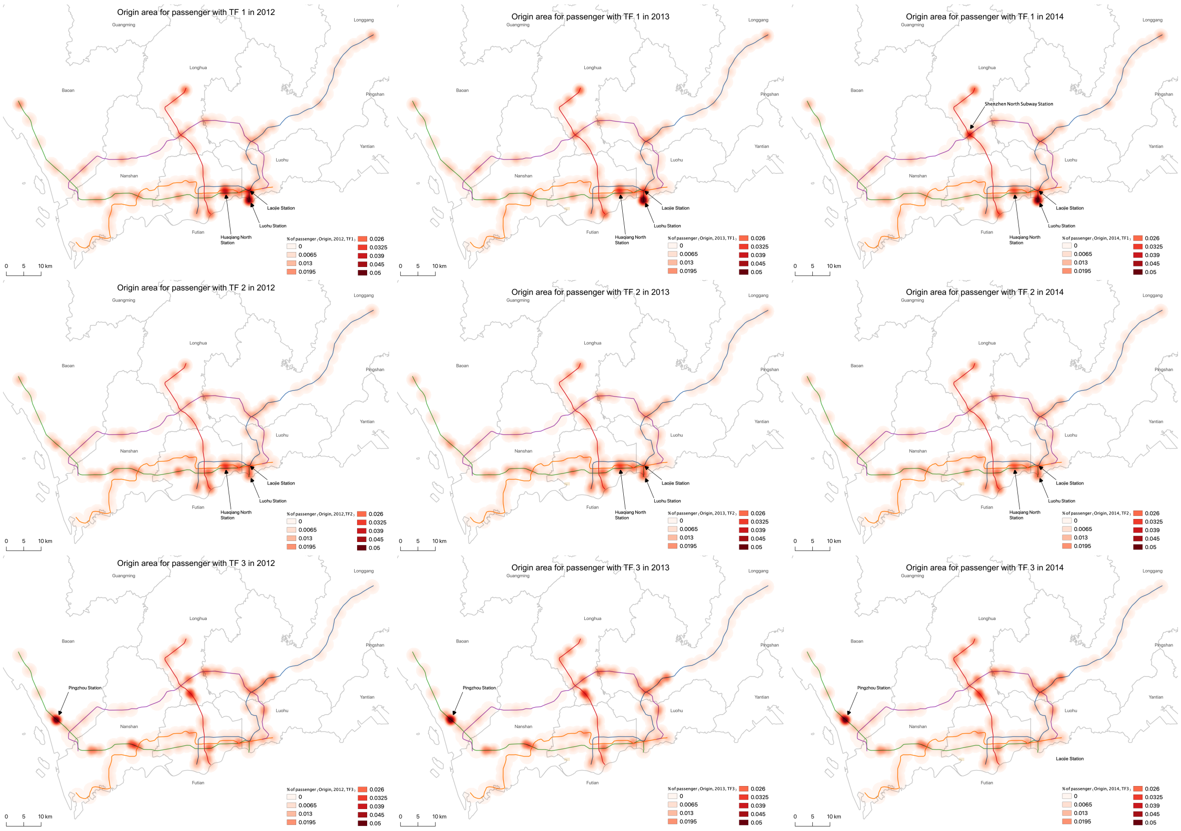
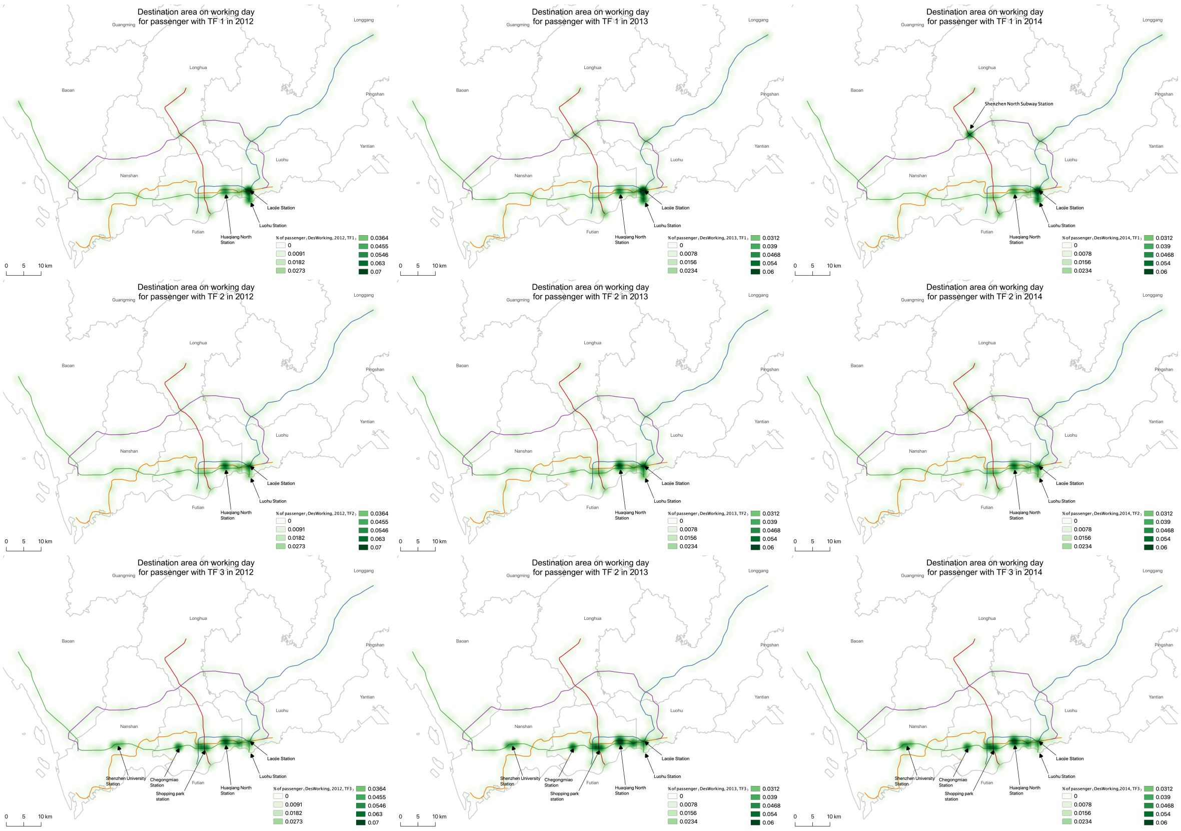

Origin area for passenger with TF 1 in 2012
Origin area for passenger with TF 1 in 2013
Origin area for passenger with TF 1 in 2014
Origin area for passenger with TF 2 in 2012
Origin area for passenger with TF 2 in 2013
Origin area for passenger with TF 2 in 2014
Origin area for passenger with TF 3 in 2012
Origin area for passenger with TF 3 in 2013
Origin area for passenger with TF 3 in 2014
Guangming
Longgang
Longhua
Baoan
Pingshan
Yantian
Luohu
Nanshan
Futian
Laojie Station
Luohu Station
Huaqiang North Station
Shenzhen North Subway Station
Pingzhou Station
% of passenger (Origin, 2012, TF1)
0
0.0065
0.013
0.0195
0.026
0.0325
0.039
0.045
0.05
0
5
10 km

Figure 9: The origin area for passengers with different trip frequency



Destination area on working day for passenger with TF 1 in 2012

Destination area on working day for passenger with TF 1 in 2013

Destination area on working day for passenger with TF 1 in 2014

Destination area on working day for passenger with TF 2 in 2012

Destination area on working day for passenger with TF 2 in 2013

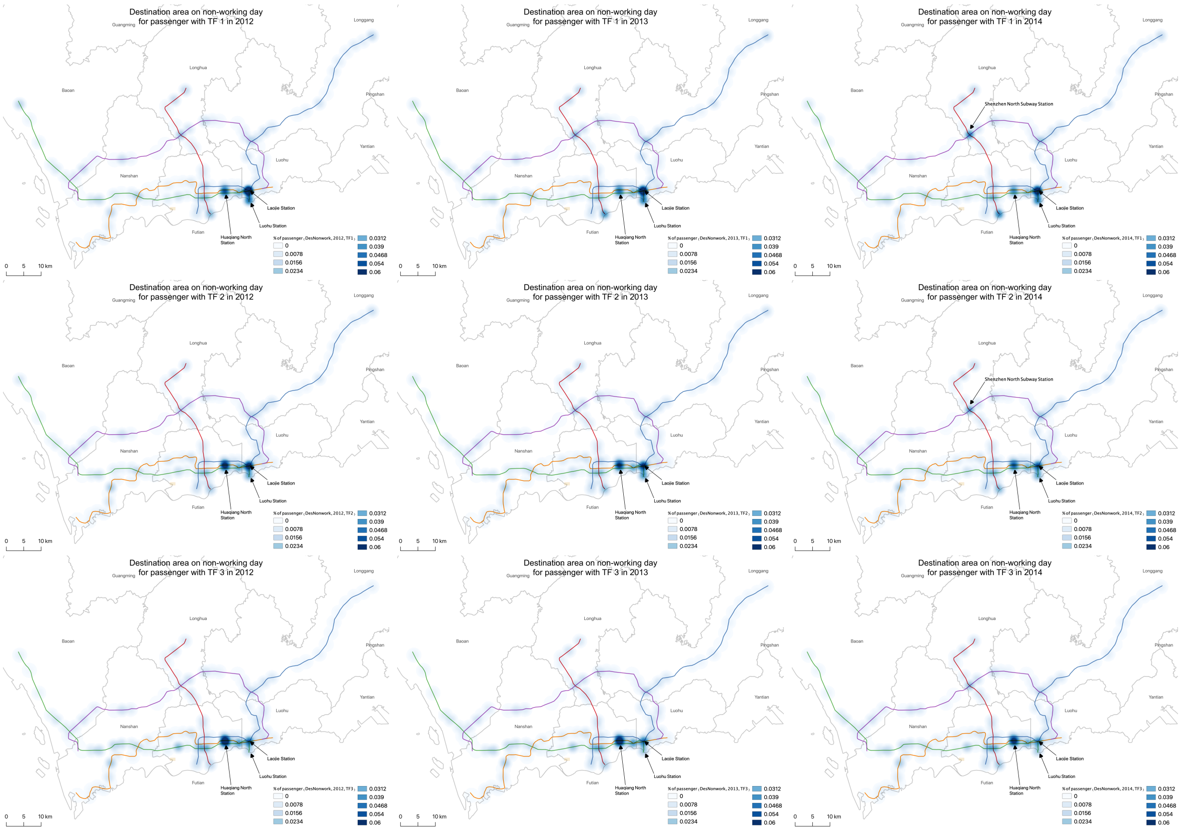
Destination area on working day for passenger with TF 2 in 2014

Destination area on working day for passenger with TF 3 in 2012

Destination area on working day for passenger with TF 2 in 2013

Destination area on working day for passenger with TF 3 in 2014

Figure 10: The destination area on working day for passengers with different trip frequency



Destination area on non-working day for passenger with TF 1 in 2012

Destination area on non-working day for passenger with TF 1 in 2013

Destination area on non-working day for passenger with TF 1 in 2014

Destination area on non-working day for passenger with TF 2 in 2012

Destination area on non-working day for passenger with TF 2 in 2013

Destination area on non-working day for passenger with TF 2 in 2014

Destination area on non-working day for passenger with TF 3 in 2012

Destination area on non-working day for passenger with TF 3 in 2013

Destination area on non-working day for passenger with TF 3 in 2014

Figure 11: The destination area on non-working day for passengers with different trip frequency

5.4 Member switching analysis

Through statistical analysis, it can be found that passengers with the same trip frequency have similar PTCs in different periods. The reason for this may be that the trip frequency of most passengers is relatively stable, so the final reaction of PTCs has not changed. Another possible reason is that the passenger's trip frequency has changed, but their PTCs will also change accordingly, becoming the PTCs corresponding to the new trip frequency. For verification, Table 6, Table 7, Table 8 counts the trip frequency for each passenger at each period.

Period 1	***Period 2***	***Period 3***	***# of passengers***	***Proportion***
TF1	TF1	TF1	225576	43.8%
TF1	TF1	TF2	83320	16.2%
TF1	TF1	TF3	12412	2.4%
TF1	TF2	TF1	79410	15.4%
TF1	TF2	TF2	72287	14.0%
TF1	TF2	TF3	13435	2.6%
TF1	TF3	TF1	7180	1.4%
TF1	TF3	TF2	10098	2.0%
TF1	TF3	TF3	11841	2.3%
	Total		515559	100%

Table 6: TF 1 in Period 1 as reference

Period 1	***Period 2***	***Period 3***	***# of passenger***	***Proportion***
TF2	TF1	TF1	85893	22.0%
TF2	TF1	TF2	50181	12.9%
TF2	TF1	TF3	8053	2.1%
TF2	TF2	TF1	71784	18.4%
TF2	TF2	TF2	105156	27.0%
TF2	TF2	TF3	20800	5.3%
TF2	TF3	TF1	9512	2.4%
TF2	TF3	TF2	18114	4.6%
TF2	TF3	TF3	20562	5.3%
	Total		390055	100.0%

Table 7: TF 2 in Period 1 as reference

Period 1	***Period 2***	***Period 3***	***# of passenger***	***Propotion***
TF3	TF1	TF1	10243	8.6%
TF3	TF1	TF2	6820	5.7%
TF3	TF1	TF3	2400	2.0%
TF3	TF2	TF1	11768	9.9%
TF3	TF2	TF2	18569	15.5%
TF3	TF2	TF3	8383	7.0%
TF3	TF3	TF1	8867	7.4%
TF3	TF3	TF2	17137	14.3%
TF3	TF3	TF3	35251	29.5%
	Total		119438	100%

Table 8: TF 3 in Period 1 as reference

As can be seen from the results of Table 6, Table 7, Table 8, taking the passenger's trip frequency in Period 1 as a reference, for TF 1, 2 and 3, the proportion of passengers whose trip frequency remains the same is the highest. But the highest ratio is just over 40%. This shows that most of the passenger with TF 1/2/3 in Period 1 have changed their trip frequency in Period 3.

That is to say, the reason why the PTCs of TF 1, 2 and 3 showed no significant change in different periods is that, after the passengers change the trip frequency, their PTCs also change accordingly. For example, a passenger in Period 1 at TF 1, his daily ridership on non-working days is greater than the working day. When his trip frequency in Period 2 becomes TF 3, he would switch to have a larger daily ridership on working day.

5.5 Summary

According to the results in this chapter, the following summary can be made:

- In the same period, the PTCs described by the four indicators show difference if the related passengers have different trip frequency.
- Statistically speaking, the PTCs related to passengers with TF 1/2/3 show no significant difference in Period 1, 2 and 3.
- Based on origin/destination area, the percentage distribution of passenger experienced changes near Shenzhen North Subway Station, and these changes involved varies among passengers with different trip frequencies.
- At different periods, a group of passengers with the same frequency of travel performed almost identical PTCs. But for this group of passengers the same trip frequency, most of the members experienced an update during each period

6 Conclusion

It can be seen from the above results that the PTCs described by the four indicators have not evolved statistically from 2012 to 2014. However, when discussing based on the origin/destination area, the distribution of passengers at individual trip frequencies varies at individual stations, and these changes are related to the passenger's trip frequency. On this basis, the trip frequency of most passengers has changed, and the PTCs involved have also changed according to the new trip frequency. Eventually, as long as the travel frequency is the same, the PTCs presented at different period are roughly the same.

In conclusion, PTCs generated by the sample data in this thesis showed a slight evolution from 2012 to 2014, and passengers involved in the sample data with different travel frequencies show different PTCs evolution. This also proves that the trip frequency can be used to find the difference in passenger's PTCs evolution. In addition, not all indicators will show change during the evolution process. In the further research, it is interesting to study the selection of indicators and sample passengers to find the more obvious evolution of PTCs

References

[1] Agard, B., Morency, C., and Trépanier, M. (2006). Mining public transport user behaviour from smart card data. *IFAC Proceedings Volumes*, 39(3):399–404.

[2] Bhaskar, A., Chung, E., et al. (2014). Passenger segmentation using smart card data. *IEEE Transactions on intelligent transportation systems*, 16(3):1537–1548.

[3] Briand, A.-S., Côme, E., Trépanier, M., and Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, 79:274 – 289.

[4] Chu, K. K. A. (2015). Two-year worth of smart card transaction data–extracting longitudinal observations for the understanding of travel behaviour. *Transportation Research Procedia*, 11:365–380.

[5] Goulet-Langlois, G., Koutsopoulos, H. N., and Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies*, 64:1–16.

[6] Hasan, S., Schneider, C. M., Ukkusuri, S. V., and González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318.

[7] Hodges, J. (1958). The significance probability of the smirnov two-sample test. *Arkiv för Matematik*, 3(5):469–486.

[8] Huang, J., Levinson, D., Wang, J., Zhou, J., and Wang, Z.-j. (2018). Tracking job and housing dynamics with smartcard data. *Proceedings of the National Academy of Sciences*, 115(50):12710–12715.

[9] Lee, S. G. and Hickman, M. (2013). Are transit trips symmetrical in time and space? evidence from the twin cities. *Transportation Research Record*, 2382(1):173–180.

[10] Ma, X., Liu, C., Wen, H., Wang, Y., and Wu, Y.-J. (2017). Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58:135–145.

[11] Ortega-Tong, M. A. (2013). *Classification of London's public transport users using smart card data.* PhD thesis, Massachusetts Institute of Technology.

[Shenzhen Metro Group Co., Ltd.] Shenzhen Metro Group Co., Ltd. Shenzhen metro introduction. `http://www.szmc.net/`. Accessed:2019-06-05.

[13] Viallard, A., Trépanier, M., and Morency, C. (2019). Assessing the evolution of transit user behavior from smart card data. *Transportation Research Record*, page 0361198119834561.

[14] Wang, Z.-j., Chen, F., Wang, B., and Huang, J.-l. (2018). Passengers' response to transit fare change: an ex post appraisal using smart card data. *Transportation*, 45(5):1559–1578.

COLOPHON

This document was typeset using LATEX. The document layout was generated using the arsclassica package by Lorenzo Pantieri, which is an adaption of the original classicthesis package from André Miede.

TUDelft