SPATIAL BIG DATA ANALYTICS OF SPATIOTEMPORAL MOBILITY CHARACTERISTICS OF THE ELDERLY

ZHICHENG SHI

PhD

The Hong Kong Polytechnical University

2020

The Hong Kong Polytechnic University

Department of Land Surveying and Geo-Informatics

SPATIAL BIG DATA ANALYTICS OF SPATIOTEMPORAL MOBILITY CHARACTERISTICS OF THE ELDERLY

ZHICHENG SHI

A Thesis Submitted in Partial Fulfilment of the Requirements for

the Degree of Doctor of Philosophy

June 2020

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

ZHICHENG SHI (Name of student)

Abstract

Many countries worldwide have ageing populations. In 2019, the size of the global population of the elderly who aged 65 years or over was 703 million. As the number of the elderly population increases, they will need more care and help. People's basic needs include clothing, food, housing and transport, and this is just as true for the elderly. The thesis focused on study the elderly behavior, which is still an open issue. Specifically, the thesis considered the spatiotemporal and mobility characteristics of the elderly and took Beijing as a study area.

Existing research based on annual traffic surveys or questionnaires on the spatial distribution of the elderly lacks near real-time prediction (e.g., next hour), large sample sizes (in the millions), precise location (latitude and longitude coordinates), and high frequency (hourly) data. Research based on real-time spatial big data, such as smart card data, can fill this research gap. The aim of this thesis was to understand the spatiotemporal and mobility characteristics of the elderly by using smart card data. This aim was achieved by realizing the following interrelated research objectives:

1. To develop a Voronoi construction method based on an integrated clustering method for region partition.
2. To study the spatial distribution characteristics and mobility behavior of the elderly. The spatial distribution characteristics include a) identifying their home locations using smart card data, b) explaining why the elderly distribution like this, and c) detecting places they visit frequently. The mobility behavior include a) travel time, b) travel distance, c) travel duration, and d) travel frequency.

A data-driven methodology was adopted for this thesis. Smart card data for Beijing, a city with a high life expectancy, were used in the analysis.

To achieve objective 1, an integrated method was developed to detect clusters in datasets with multiple densities and shapes features. Two improvements were made to the classic clustering methods: a) cluster number was estimated automatically, and b) only one parameter was required. With these improvements, multiple densities and shapes of clusters could be detected effectively.

The clustering method was also scalable for different kinds of dataset.

To achieve objective 2, three targets were set to examine and determine the spatial distribution patterns of the elderly by means of smart card data. First, the spatial distribution of the elderly population in a city was analyzed using the Voronoi diagram which is based on an integrated clustering method. The proposed method can efficiently detect clusters with multiple densities and shape and accurate to present the spatial distribution of elderly. Second, the spatial distribution pattern of the elderly was measured and explained by a newly proposed model of PoI-based elderly livability index computed based on weighted factors including restaurants, parks, hospitals, shops and bus stops. Third, the spatial connectivity between regions in which the elderly travels was used to describe where the elderly frequently travels. A quasi-gravity model was developed to reveal the relationship between spatial connectivity and the PoI-based elderly livability index.

Three important findings were yielded: a) the spatial distribution of the elderly's home locations shows clear clustering characteristics; b) the spatial distribution of the elderly has a strong relationship with public service facilities, such as restaurants and hospitals; and c) the connectivity of pairs of regions is related to the distribution of public facilities in the connected regions.

To understand the mobility behavior of the elderly, two methods were adopted: (a) quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent presentation of such behavior graphically, and (b) discovery of the distribution functions of the travel characteristics, both by function curve fitting and by testing the goodness of fit of the identified distribution functions.

A number of important findings have yielded on elderly mobility behaviors in megacities: (a) most of the elderly's travels are approximately 1 km and the median distance is 4 km, which is shorter than the adults, and their travel distance follows an exponential function, unlike the travel distance distribution of adults, which follows a Gaussian function; (b) most of the elderly travel for 4 minutes, which is half the time of the adults' travels, and travel time follows a Gaussian distribution; (c) the

elderly's travel departure time has a morning peak at 9:00 am (compared with 8:00 am for adults) and no clear peak in the afternoon; and (d) most of the elderly travel once per day, which is the same as the adults.

The significance of this thesis lies in the series of new analytics methods developed and the comprehensive findings regarding the spatial distribution patterns and mobility behavior of the elderly in megacities. These findings add new knowledge to the field. The new methods could be widely applied in urban planning, management and services for the aging population.

Acknowledgements

First, I would like to express my sincere and endless gratitude to my chief supervisor, Dr. Lilian PUN for her constant support, invaluable advice and patient guidance. Without her help, I would have given up and it would be possible to complete the thesis. Dr. Lilian PUN is a knowledgeable advisor and patient mentor. She is not just an excellent researched, but also a patient teacher. I have gotten many benefits from her under her guidance. She always provide many valuable insights and feedback about my research problems.

I am deeply grateful to the academic staff and administrative staff members in the Department of Land Survey and Geo-informatics for their support during my Ph.D. study period. They kindly answer my research questions and patient provide constructive suggestions.

I am very thankful for my friends. We had a wonderful time together during my study period. They always encourage me and help me to cheer up when I encountered difficulties. They also provide valuable suggestions and opinions about my research.

I would like to thank the financial support from The Hong Kong Polytechnic University. Without the great support, I would not have the chance to study in Hong Kong.

Last but not least, I would like to deeply thank my family for supporting me throughout the study period and my life in general.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

As the global elderly population continues to increase, population ageing is becoming a challenging issue for most countries in the world. The global population reached 7.7 billion in the middle of 2019 and is expected to reach 9.7 billion by 2050 (United Nations, 2019a). Although the global population is increasing, the rate of increase is falling. By 2015, the growth rate had fallen below 1.1 percent and it is projected to reduce further. Some countries are experiencing a decrease in total population size.

The pattern is different for the elderly population. In 2019, the size of the global population aged 65 years or over was 703 million (United Nations, 2020). By 2050, the number is estimated to reach 1.5 billion. As the size of the global elderly population grows, it is becoming an increasingly large proportion of the overall population. The proportion of the global population aged 65 or over increased from 6 percent in 1990 to 9 percent in 2019 and is estimated to rise to 16 percent in 2050. It is estimated that by 2050, one in six people will be aged 65 years or over.

As the world economy has developed, urbanization has increased in contemporary societies. Population migration from rural to urban areas is an important characteristic of the urbanization process. In 1950, 30 percent of the global population was living in urban areas. In 2018, the proportion was 55 percent and it is projected to be 68 percent in 2050 (United Nations, 2019b). More and more megacities with a population of more than 10 million are continuing to form. There were 20 megacities in Asia in 2018. It is thus expected that the elderly population in megacities will continue to increase.

Therefore, it is essential that governments make innovative policies and improve the quality of public services to solve the problems faced by the elderly in megacities, such as housing, traveling

and health care. Understanding the spatial and temporal characteristics of the elderly by using smart card data is a research direction.

1.2 Related works

1.2.1 Clustering methods

Clustering methods play more important role in the era of big data. They are widely used in many research domains. It aims to group events according to neighboring occurrence and/or similar attributes. Most clustering algorithms should measure the distance between each pair. Various distance functions are adopted in the clustering methods, such as the Euclidean and Manhattan distance functions. A famous application of clustering occurred in 1854, when Dr John Snow found that clusters of cholera cases occurred around a public water pump, which was the source of the spread of cholera.

Different domains include various types of objects such as location information and customer consumption (Bradlow et al., 2017). Human mobility can be detected from location information from traveling (Zheng, 2015). For example, commuting is one important pattern that can be determined from smart card data. Homes and workplaces represent clear clustering characteristics. Shops can improve their marketing strategy by understanding similar consumption characteristics. To measure similarity, distance functions are mostly used to calculate the distance between objects and grouping the objects on the basis of the results to form clusters is the core process. Many clustering methods have been proposed in recent decades.

Conventional clustering methods can be divided into four categories: partitioning, hierarchical, density-based, and grid-based methods (Han et al., 2011). Partitioning methods divide whole objects into a specified number of datasets. One popular method is k-means (Raykov et al., 2016). The core idea is to divide the dataset into k clusters, and the k value is normally defined manually. For each object, the nearest cluster centroid is identified to form clusters. The centroids of new clusters are

calculated until they do not change to achieve the best clustering results. K-medoids (Park & Jun, 2009) is a similar method to k-means. The difference is the selection of initial centroids. Both methods are easy to implement and highly efficient, but the cluster results are affected by the selection of the initial centroids.

For hierarchical clustering methods, clusters can be formed step by step from the top down or bottom up. Objects are merged with others based on the shortest distance to form clusters until certain conditions are satisfied. The bottom-up method is called agglomerative hierarchical clustering. In the opposite process, divisive hierarchical clustering, the whole dataset is partitioned into a number of clusters until the conditions are satisfied. BIRCH adopts the clustering concept to form a clustering tree to conduct the cluster process (T. Zhang et al., 1996). The cluster information is stored in tree form. This method has better clustering quality and the ability to deal with large datasets. CURE (Guha et al., 1998) and Chameleon (Karypis et al., 1999) are two other hierarchical clustering methods. With CURE, the nearest objects are merged until the target is achieved. Instead of using one object or a centroid to serve as a cluster, several objects are selected to represent the cluster by multiple a shrinkage factor. Chameleon is a two-stage clustering method; the nearest points are merged to form small clusters, and small clusters with a high value of relative interconnectivity and relative closeness are then merged.

Density-based methods can locate clusters of arbitrary shape. Three representative methods are DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999) and DENCLUE (Hinneburg & Gabriel, 2007). DBSCAN requires a minimum number of neighborhood objects, and the maximum radii of neighbors are predefined by the user. Objects are divided into three categories: core objects, reachable objects, and noise. Each cluster is formed by core objects and reachable objects until all objects are assigned to one of the three categories. However, the parameter settings are mostly determined according to the user's experience to choose distance parameter. To overcome this problem, OPTICS does not form clusters. Instead, it generates a cluster ordering to represents the cluster results in a graph. The correct number of clusters cannot be calculated by this method. DENCLU adopts a density distribution such as Gaussian kernel to estimate the density to investigate

clusters of objects. It can reduce the influence of noise via Gaussian or different kernel functions. The discovery of clusters with a non-spherical shape is one of its main advantages. These methods use a data-driven focus to partition the dataset into many clusters.

Grid-based methods use a space-driven focus to separate a space into cells and assign objects to them. STING (W. Wang et al., 1997) and CLIQUE (Uncu et al., 2006) are two representative examples of grid-based clustering methods. STING is a multiresolution clustering method that partitions a space into cells with a hierarchical structure. The number of cells gradually increases from the high to low level. The size of each cell in the high level is formed by a number of cells from the level just below. The quality of the cluster results depends on the appropriate size of the cells. If a cell is too large or too small, the accuracy of the clusters may be affected. CLIQUE uses a grid-based method and also considers the density of objects and multiple dimensions. It partitions each dimension into different non-overlapping levels and assigns whole objects into cells. A cell is identified as dense when the number of objects exceeds a density threshold. After the first dense cell is located, the neighboring cells are merged if they are also dense until no further high-density cells can be found. The process is reiterated until all cells are marked as high-density or low-density.

Clustering is also a high-performance tool for detecting hot spot patterns in spatiotemporal data analysis (Han et al., 2011). Spatiotemporal data analysis methods can be classified into six categories—clustering, prediction, change detection, frequent pattern mining, anomaly detection, and relationship mining (Atluri et al., 2017). Clustering has been used in many applications with spatiotemporal data (T. Cheng et al., 2014).

1.2.2 Studies on the mobility of human and the elderly

The rapid development of human mobility analytics in recent decades has been stimulated by two factors: the ability to collect large quantities of data from a variety of sources, such as transportation (Zheng, 2015) and social media (Tang et al., 2014) , and the proposal and application of many new data mining technologies in many areas. Big data and new technological methods have laid the

foundation for a greater understanding of human mobility behavior; in particular, it has been noted that such behavior involves spatial and temporal regularity, particularly regarding individual mobility (Gonzalez et al., 2008). Citizens usually require a high frequency of round trips, such as traveling between home and the workplace. It is possible to use such travels as models to predict sequences of common human mobility needs (Song, Koren, et al., 2010; Song, Qu, et al., 2010).

Two data source categories are generally used for analyses of human mobility: questionnaire surveys and sensor data. Surveys are used mainly to investigate factors that affect travel behavior. Survey data focus on detailed trip and traveler information, such as age, gender, and job nature. In Korea, household travel survey data have been used to investigate whether personal and household characteristics affect potential travel, and the results showed that the main factors that affect travel vary significantly among age groups (Hahn et al., 2016). In Hong Kong, interview survey data and Hong Kong travel characteristics survey data have been used to study transport mobility, and it was found that the elderly's travel mobility characteristics differ from those of younger people (He et al., 2018; Wong et al., 2018). A possible reason lies in the stable maturity of Hong Kong's public transportation systems. By analyzing and comparing the elderly and young people who taking public transportation, the distinctions of their travel characteristics during different time periods were analyzed (Shao et al., 2019).

Domestic and foreign studies have shown differences in the elderly's travel behavior, especially in terms of travel modes. Social economies, urban structures, and policy-making have led to different travel modes in different countries. For instance, in the United States, Europe, and Australia, driving by private car plays a dominant role in the elderly's traveling (Böcker et al., 2017; Boschmann & Brady, 2013; Cui et al., 2017; Truong & Somenahalli, 2015). However, some studies have shown that the percentage of the elderly who travel by car is reducing (McDonald, 2008). Even though driving a private vehicle is the most popular transportation mode, the study showed that around 36.6% of the elderly lack for travel modes (Van den Berg et al., 2011). The situation is different in other countries. Beijing is a highly dense city in China. To avoid traffic jams on the limited road network, walking is an ideal mode preferred by most the elderly, as is also the case in Changchun (X. Hu et

al., 2013; Liu et al., 2017). In Seoul, walking is the most common transport mode, and travel by car ranks second. As age increases, the difference between them becomes more obvious (Choo et al., 2016). Hong Kong has a developed public transportation system that provides satisfactory barrier-free facilities and services for the elderly. Taking the bus is their first choice (Szeto et al., 2017). In addition to regional differences, economic and weather conditions such as income and temperature have strong relationships with the selection of travel mode (Böcker & Thorsson, 2014; Kim, 2011; Moniruzzaman et al., 2015; Páez et al., 2007).

Travel distance and frequency are both important characteristics that can determine and therefore illustrate the efficiency of human mobility, especially when studying the activity of the elderly. In general, the elderly with physical or mental impairments caused by normal physical deterioration due to age do not enjoy the prospect of frequent travel, particularly when using public transport. People with reduced physical ability and stamina generally travel less frequently and over shorter distances (Collia et al., 2003; X. Hu et al., 2013). With improvements in life quality and the use of medical technologies, the gap between the elderly and the adults is not as large as once perceived. Walking accessibility has been used to investigate the elderly access to recreation amenities in various areas. Results of studies have indicated that different districts have an uneven distribution of accessible recreation places for the elderly (L. Cheng, Caset, et al., 2019). The built environment influences the travel behavior in terms of frequency and time expenditure. Certain social and cultural factors make the travel behavior of the elderly very distinctive (L. Cheng, Chen, et al., 2019). In particular, the distribution of public facilities such as public transportation, vegetable markets, and parks with recreation activities affect the travel behavior of the elderly (Feng, 2017). Many sociodemographic variables affect travel behavior (Yang, 2018). The distribution of public facilities and satisfaction with public transport services are both research problems. According to the survey results (Wong et al., 2017), seat availability has the lowest satisfaction level among the service aspects. This can suggest to government that enhancing service quality would support the elderly's travel. In Harbin, a customer satisfaction index was proposed to evaluate passenger satisfaction with the bus transport services. Several indicators such as reliability, time schedule, and security were used to perceive service quality (Yuan et al., 2019). Y. Zhang et al. (2019) used the Travel Survey

of Beijing Inhabitants to describe detailed travel information such as departure times, travel purpose and travel modes. At the same time, they investigated the free bus program has little effect for the elderly mode choice.

The connectivity between places can be detected from human travel by means of different modes. because of similarly individual travels happened in the certain space and time (Gonzalez et al., 2008). Many data sources can be used such as smart card and mobile phone data (Jiang et al., 2017; Long & Thill, 2015; Sun et al., 2012; Zhong et al., 2016). By exploring these data can help to understand the connectivity between places for helping urban structure and planning. They are valuable information can support to improve city competition and lay foundation to future development (Barbosa et al., 2018; Zhong et al., 2016; Zou et al., 2018). Elderly population distribution and connectivity have been studied based on the different data sources. The different region areas demonstrate both similarities and different patterns. Because of the diversities of geography, population and culture, the elderly population distribution and connectivity in different countries and regions quite markedly. For instance, in Europe, America and Australia, the private car is the main transportation tool (Boschmann & Brady, 2013; Cui et al., 2017; Truong & Somenahalli, 2015). These countries are sparsely populated areas which is suited to develop the highway transportation. It leads to elderly living dispersedly without high density as well as public facilities distribution. Mostly, their travels rely on private cars, and thus they can have longer connectivity of distance of different regions. But in China, the situation is opposite. Most of the elderly live in the city center where a high density of public facilities generally exists.

Regarding the latter point, a clear clustering distribution of elderly is seen in the above areas. With economic development and population growth, private cars are not the only major transportation tool used by the elderly. Walking, bicycles and public transport are population travel modes in China (X. Hu et al., 2013; Liu et al., 2017; Shao et al., 2019; Y. Zhang et al., 2019). The connectivity of distance of different places is obviously shorter than many developed countries or regions. For example, Hong Kong has a highly developed and efficiently developed public transportation system (Yang, 2018) together with complete social welfare to encourage elderly in social activities. Public

transportation facilities in most areas of Hong Kong have proved satisfactory in that quality service is provided (Szeto et al., 2017). The connectivity of short distance is more than long distance. In summary, elderly population distribution and connectivity of different places is influenced by urban structure and public services facilities distribution (Guo et al., 2019) such as public transportation.

Elderly population distribution and spatial connectivity are not only related to travel modes, but also related to activity areas and service facilities (L. Cheng, Chen, et al., 2019). Many service facilities such as shops, hospitals and bus stops are influential lifestyle factors (Ahern & Hine, 2012; Titheridge et al., 2009), such factors can affect the connectivity of various regions. For example, the elderly like to take activities in local areas if there is not good public transport services (Findlay et al., 2001; Goins et al., 2005; Lin et al., 2014; Yuan et al., 2019). If there is a strong connectivity between two places, the value of the total number and frequency are large as the elderly are more easily enabled to have round trip between them(Plazinić & Jović, 2018). The study results show that public service facilities have a strong impact on connectivity between different places.

1.3 Aim, scope and objectives

1.3.1 Aim

The aim of this thesis was to understand the spatiotemporal and mobility characteristics of the elderly using spatial big data from smart cards data.

1.3.2 Scope

The scope of the thesis was to develop methods of collecting and analyzing the spatial distribution patterns and mobility behavior of the elderly using spatial big data. The methodology emphasizes data-driven methods from the “fourth paradigm of science” in the era of big data. Clustering methods in spatial data mining and quantitative analysis were used to investigate. The spatial

distribution of the home locations of the elderly, the reasons why the elderly distribution like this, and which places they visited frequently. The mobility data investigated included travel distance, travel duration, travel time and travel frequency when using the bus, the most popular transport mode for the elderly. Smart card data and points of interest (PoI) data were used in this thesis. PoI data provide detailed information on facilities used by the elderly and smart card data provide travel records at the million-level sample size with near real-time, high-accuracy location information. The thesis area was Beijing, which is not only one of the largest cities in the world, but also a city with high life expectancy.

1.3.3 Objectives

The aim of this thesis, which was to understand the spatial and mobility characteristics of the elderly using spatial big data, was achieved by realizing the following interrelated research objectives:

1. Objective 1, to develop a Voronoi diagram construction method based on an integrated clustering method for region partition. The integrated clustering method can detect clusters with multiple densities and shape distributions.
2. Objective 2, to study the spatial distribution characteristics and mobility behavior of the elderly. The spatial distribution characteristics include a) identifying their home locations using smart card data, b) explaining why the elderly distribution like this, and c) detecting places they visit frequently. The mobility behavior include a) travel time, b) travel distance, c) travel duration, and d) travel frequency.

To achieve objective 1, an integrated method was developed to detect clusters in datasets with multiple densities and shapes. Two improvements were made to the classic clustering methods: a) cluster number was estimated automatically, and b) only one parameter was required. With these improvements, multiple densities and shapes of clusters could be detected effectively. The clustering method was also scalable for different kinds of dataset.

To achieve objective 2, three targets were set to examine and determine the spatial distribution patterns of the elderly using smart card data. First, the spatial distribution of the elderly population in a city was analyzed using the Voronoi diagram which is based on the integrated clustering method. Second, the spatial distribution pattern of the elderly was measured and explained by a newly proposed model of PoI-based elderly livability index that is computed based on weighted factors including restaurants, parks, hospitals, shops and bus stops. Third, the spatial connectivity between regions in which the elderly live was used to describe where the elderly frequently travels to. A quasi-gravity model was developed to reveal the relationship between spatial connectivity and the PoI-based elderly livability index. Two methods were adopted to study mobility behavior: (a) quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent presentation of such behavior graphically, and (b) the discovery of the distribution functions of the travel characteristics, both by function curve fitting and by testing the goodness of fit of the identified distribution functions.

1.4 Research problems

In the thesis, a data-driven method was used to analyze spatial big data. Clustering is an important method widely used in such analysis. However, there are two problems associated with existing methods: (a) how to detect clusters with multiple densities and shapes, and (b) how to reduce the number of parameters that must be predefined before clustering. A new clustering method was needed to overcome these problems.

Four aspects of the elderly's mobility behavior in megacities were identified as the research problem, as follows. a) what is the travel distance of the elderly? b) what are the travel departure and arrival times of the elderly? c) what is the travel frequency of the elderly? d) what is the travel duration of the elderly? Understanding these mobility behaviors is essential for transport planning, management, and provision of services for the elderly.

To study the spatial characteristics of the elderly, the following questions were considered: a) where

do they live? b) why the elderly distribution like this? c) where do they go frequently? A series of methods were developed to explore the distribution patterns of the elderly population in Beijing using smart card data and PoI distribution data. The research findings provide an important scientific basis for policymaking and urban planning.

1.5 Research methodology

1.5.1 Methodology

The data-driven methodology adopted for this thesis included (a) an integrated method to detect clusters in datasets with multiple densities and shapes; (b) quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent presentation of such behavior graphically, and (c) discovery of the distribution function of the travel characteristics, both by function curve fitting and by testing the goodness of fit of the identified distribution functions. These methods are described in detail below.

Voronoi construction based on an integrated spatial clustering method

Voronoi construction based on an integrated spatial clustering method was developed to detect clusters in datasets with multiple densities and shapes to construct Voronoi diagram for city partition. Two improvements were made. First, the number of clusters did not need to be pre-assigned. Second, only one parameter was needed. With these improvements, objects with multiple densities and shapes of clusters could be detected effectively. The clustering method could also be applied to different kinds of dataset. The region partition method is suitable than administrative regions.

A framework for analyzing the mobility behavior of the elderly

A framework composed of data capture, data preprocessing (including data cleaning), and mobility analytics was developed based on the four identified features of interest (travel distance, departure and arrival time, travel frequency, and travel duration). Two data-driven methods were used for the analysis: a) quantitative analysis of spatiotemporal travel behavior by estimating the parameters of

travel patterns and the subsequent presentation of such behavior graphically, and b) discovery of the distribution functions of the travel characteristics.

A data-driven framework for analyzing the spatial distribution of the elderly

A framework was designed to systematically answer the following three questions: a) where do the elderly live? b) why the distribution like this? c) where do they go frequently? The framework includes the following key methods: (a) a method of identifying the home locations of the elderly based on swiping card information (four assumptions were made to identify the home location), (b) a Voronoi diagram-based method of partitioning a city into regions that is suitable to present distribution of the elderly, and (c) Voronoi diagram construction based on an integrated clustering method is proposed for region partition.

1.5.2 Research hypotheses

When analyzing the spatial distribution pattern of the elderly, their home locations can be estimated from smart card data. The following four hypotheses are were made in this thesis.

Hypotheses (a) Home location is close to the most frequently visited bus stop

It is possible to estimate the home locations of the elderly using the locations of their most frequently visited bus stops. Normally, for a commuter, home and work are the two most frequent places visited in the daily itinerary of trips. As the elderly do not have a regular workplace, their home location is a high-frequency travel origin and destination.

Hypotheses (b) The bus stop of first departure or latest arrival during an elderly person's daily travel is very likely to be the home location.

For each elderly person, there is a high possibility that the earliest departure stop is the home location. The elderly is likely to start their day's travel from home. When they end their day's travel, they are likely to return home.

Hypotheses (c) The bus stop of first departure for all recorded days has the highest probability of being the home location.

Based on assumptions 2 and 3, frequency and time of day are the two main factors used to identify bus stops regarded as the home location. The stop that the elderly use most frequently can be regarded as their home location. Most elderly people are retired and do not have a fixed place for traveling to, such as a workplace. The most frequently visited place is most likely to be their home location.

Hypotheses (d) One elderly person owns one smart card mainly for his/her trips.

For reasons of rationality and economy, the elderly usually using only one smart card for traveling. It is very rare that one elderly person has more than one smart card. Each smart card therefore represents one elderly person.

1.6 Structure of the thesis

1.6.1 Thesis framework

The overall framework of this thesis is composed of a series of the interrelated chapters. The logical flow of the thesis is illustrated by the Figure 1.1.

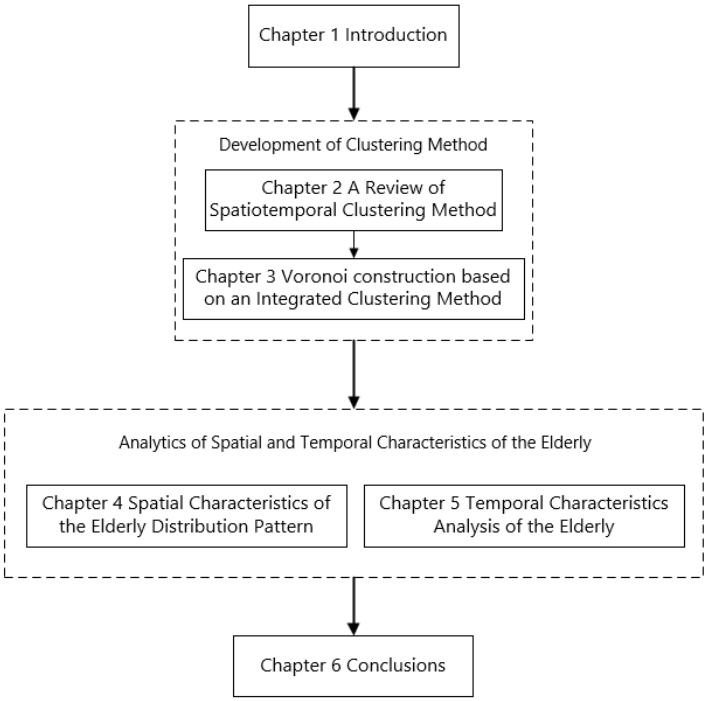


Figure 1. 1 The Structure of the thesis

The thesis mainly includes four parts: introduction, Voronoi construction based on the integrated clustering method, analysis of spatial and mobility characteristics, and conclusion. Introduction describe the thesis background of the elderly problem and scope of the proposed research. The clustering method part covers two chapters: a review of the existing cluster methods in Chapter 2, and Voronoi construction based on an integrated clustering method from this research in Chapter 3. In the part on analytics of the spatial and mobility characteristics of the elderly, Chapter 4 is on spatial distribution pattern of the elderly where the Voronoi construction method based on the integrated clustering method is used together other methods, and Chapter 5 is on temporal characteristics of the elderly. By Chapter 4 and Chapter 5, a full picture on spatial and temporal characteristics of the elderly is revealed based on smart card big data analytics. Finally, a conclusion is drawn to summarize the whole research in Chapter 6.

1.6.2 Thesis structure

Short descriptions of each chapter are provided below:

- Chapter 1: This chapter presents the background, related work, aim, scope, objectives, research methodology, and thesis structure.
- Chapter 2: A review of different spatiotemporal clustering methods is presented in this chapter. Examining the development of spatiotemporal data analysis methods can uncover potentially interesting and useful information. Due to the complexity of spatiotemporal data and the diversity of objectives, a number of spatiotemporal analysis methods exist, including but not limited to clustering, prediction, and change detection. As one of the most important methods, clustering has been widely used in many applications. It is a process of grouping data with similar spatial attributes, temporal attributes, or both, from which many significant events and regular phenomena can be discovered. Clustering methods can be divided into two categories: hypothesis testing-based methods and partition-based methods. The former mainly use a probability model and statistical hypothesis testing to find significant clusters. In general, the null hypothesis is that the distribution of events is random; if it is rejected, a cluster can be formed. Most partition-based clustering methods utilize distance functions to compute the closeness of events to distinguish cluster and noise.
- Chapter 3: In this chapter, Voronoi construction based on an integrated clustering method is described. Existing clustering methods cannot easily detect accurate cluster numbers from datasets with multiple densities and shapes. This method is proposed based on two improvements to classic clustering methods. First, it is currently difficult for users to correctly estimate cluster number. The new method automates this estimate. Second, many clustering methods require more than one parameter. Similar to cluster number, the values of parameters must be decided based on the experience of users. This method requires only one parameter. These improvements mean that clusters of multiple densities and shapes can be detected. Even when the dataset includes high and low densities and clusters of various shapes, the method is able to detect them with high efficiency. By using big data analytics, the method can explore the potential clusters in different kinds of dataset. Then the center points of clusters are used as the seed point to construct the Voronoi diagram for partitioning region. Compared with administrative

regions, this partition method is suitable for the different kinds of data.

- Chapter 4: The distribution of the elderly population and connectivity between different regions are explored, and Voronoi construction method based on the integrated clustering method is used in this chapter. First, the spatial distribution of the elderly population in Beijing is analyzed and presented in a Voronoi diagram. The integrated clustering method is used to cluster bus stops based on the flow of elderly passengers at stops and a Voronoi diagram is constructed based on the cluster centers with the aim of partitioning the city. Four assumptions are made to identify the home locations of the elderly and these are presented in the Voronoi diagram. Second, the spatial distribution pattern of the elderly is measured and explained by a newly proposed "PoI-based elderly livability index" that is computed based on urban facilities that are assigned different weights. The five identified factors are restaurants, parks, hospitals, shops, and bus stops. The numbers and spatial distributions of the five factors for each region are computed using PoI data. Third, the spatial connectivity between the elderly living regions is used to describe where the elderly frequently travels. A quasi-gravity model is developed to reveal the relationship between the connectivity and the PoI-based elderly livability index of the elderly.
- Chapter 5: An analysis of temporal characteristics of the elderly is presented in this chapter. The travel behavior of the elderly in megacities is still an open issue and was thus chosen as the focus of this research. It is essential that such data-driven analysis is based on spatial big data that can provide travel behavior analysis in near real time (next hour or day), with a large sample size (at the 100,000s or millions level), with precise location (latitude and longitude coordinates), and with high frequency (hourly). The behavior of elderly travelers is analyzed through four spatiotemporal features: departure and arrival time, travel distance, duration, and frequency. The data for the elderly are compared with those for the adult group. Two analytical methods are adopted: (a) the quantitative analysis of spatiotemporal travel behavior based on estimating the parameters of travel patterns and the subsequent presentation of such behavior graphically, and (b) the discovery of the distribution function of the travel characteristics,

both by function curve fitting and by testing the goodness of fit of the identified distribution functions.

- Chapter 6: A summary of the thesis is provided in this chapter. In short, the thesis systematically investigates the spatiotemporal characteristics of the elderly using smart card data to reveal the mobility behavior characteristics and spatial distribution characteristics of the elderly. The data-driven methods used can be applied to the study of various problems facing the elderly. Suggestions for future research are made.

1.7 Significance of the thesis

The significance of this thesis lies in the series of methods developed and the findings, which aid comprehensive understanding of the spatial distribution patterns and mobility behavior of the elderly in megacities. These findings add new knowledge to the field and the new methods can be widely used for urban planning, management and service provision for the ageing population.

For complex data sources, the integrated clustering method can handle different types of dataset that include clusters with multiple densities and various shapes. The travel distance characteristics can be used to inform policies for short-distance bus services for the elderly (such as community bus service) to optimize the service. The travel time, duration and frequency can help transport policymakers and management to plan timetables to reduce the waiting time and traffic pressure.

Chapter 2 Spatiotemporal Data Clustering: A Survey of Methods

This chapter reviewed some representative clustering methods, especially for the spatiotemporal clustering methods, most of which are extended from spatial clustering. These methods are broadly divided into hypothesis testing-based methods and partitional clustering methods that have been applied differently research domains.

2.1 Introduction

Large-scale data mining brings new opportunities and challenges for discovering hidden valuable information from enormous data sets. In particular, with the rapid development of positioning technologies as well as the emergence of a large number of positioning devices, a vast amount of data could be easily collected from different sources. These sources could come from broad domains, including government documentary and decades of collected data, transportation (Zheng, 2015), and social media (Tang et al., 2014). For example, governments conduct censuses and own large datasets containing information about population change, human movement, and economic characteristics during different periods for planning and policy making. Many floating cars such as taxi and truck installing GPS receivers can monitor running state and record spatial and temporal information every second. Social media like Facebook and Twitter can post users' experiences at a given place and time.

All this spatiotemporal information is useful for pattern analysis in space and time. Space can be represented by an address, geographical coordinates of latitudes and longitude, or local (X, Y) coordinates. Time can be shown by year, month, and day and sometimes as detailed as hour, minute, or second. Spatiotemporal (ST) data types can be divided into five categories containing events, geo-referenced variables, geo-referenced time series, moving points, and trajectories (Kisilevich et al., 2009) (Figure 2.1). The collected datasets, regardless of if they are in tabular or graphical forms, are often too complex to be understood. An efficient spatiotemporal analysis method is important to

mine meaningful patterns for better understanding or visualization (Shekhar et al., 2011).

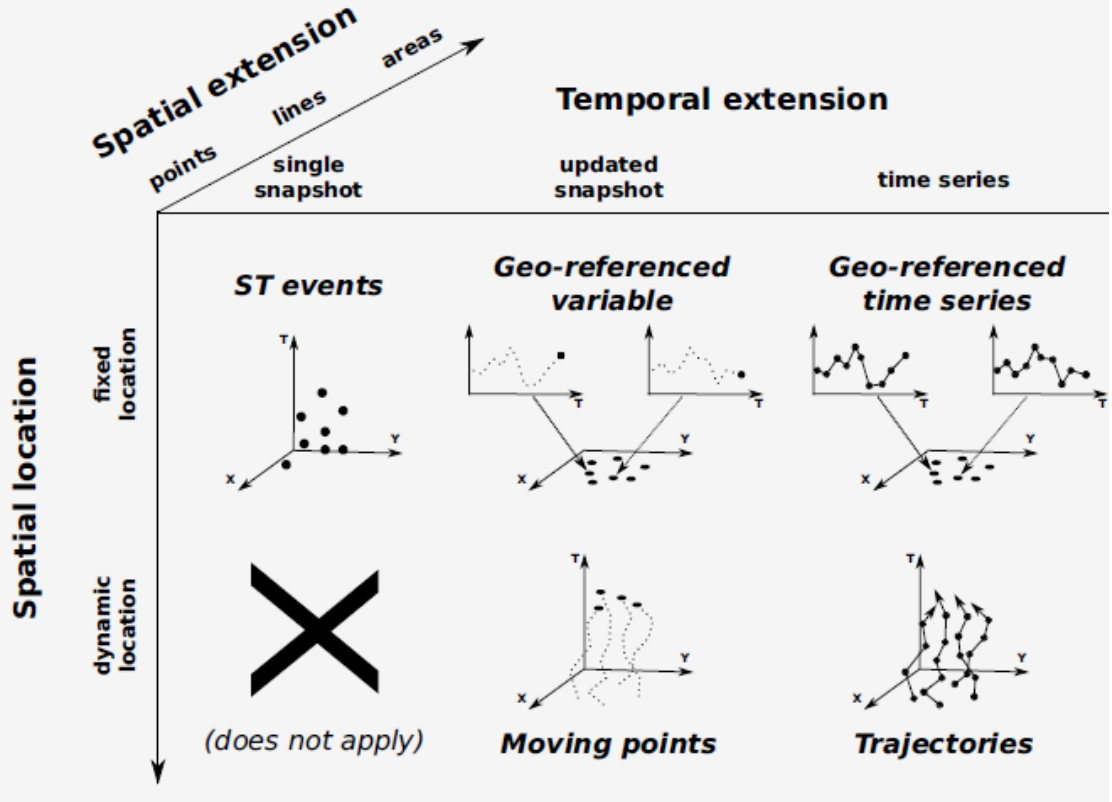


Figure 2. 1 Context for spatiotemporal (ST) clustering (source: Kisilevich et al. (2009))

In some cases, spatiotemporal clustering methods are not all that different from two-dimensional spatial clustering (Ankerst et al., 1999; Pei et al., 2009; T. Zhang et al., 1996). Figure 2.2 shows the procedure of clustering. For raw spatiotemporal data, the first step is cleaning and reorganization. Incorrect and missing data should be identified and deleted before applying an appropriate clustering algorithm. However, different parameters can affect the clustering results. It is necessary to adjust parameters for a better understanding of cluster results and interpreting potential information.

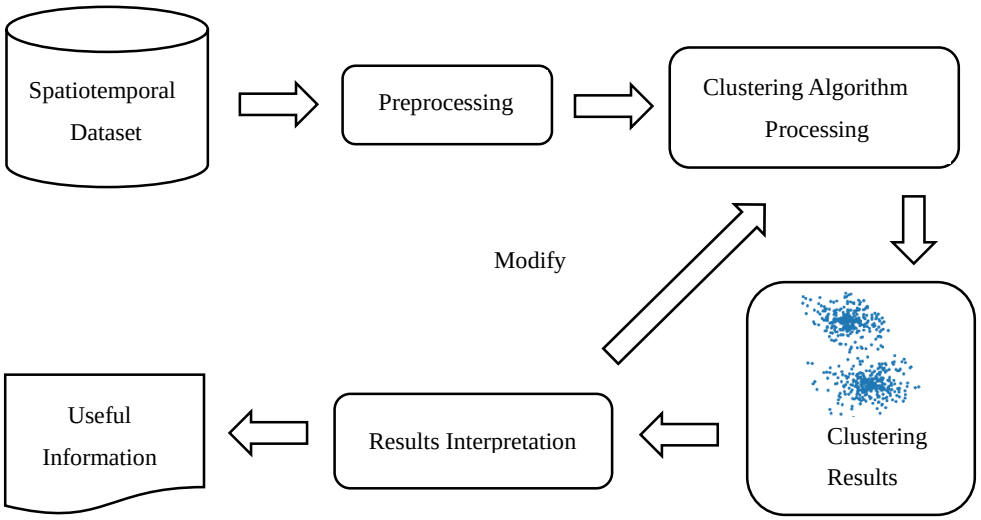


Figure 2. 2 Procedure of spatiotemporal clustering

In this chapter, we only focus on the clustering methods of the events ST data type. Some representative ST clustering methods are reviewed, most of which are extended from spatial clustering. In our view, these could be divided into two categories, the hypothesis testing-based methods and the partition-based methods. The former one mainly uses a probability model and statistical hypothesis testing to find significant clusters. In general, the null hypothesis is that the distribution of events is random; if it is rejected, a cluster could be formed. The partitional clustering methods mostly utilize distance functions to compute the closeness of events to distinguish cluster and noise. Some popular spatiotemporal clustering methods are introduced in the following sections. This will help to understand the evolution of techniques in the past decades and explore future research trends.

2.2 From spatial to spatiotemporal clustering

There is no clear definition of clustering (Rokach & Maimon, 2005; Saxena et al., 2017) and different categories have overlap such that an algorithm could contain more than one feature of categories. The major difference between spatial and ST clustering is the ‘time’ element, which is treated as either another dimension or an attribute. By space, it can be at least 2-dimensional (X,Y)

or 3-dimensional (X, Y, Z) in which events or attributes are clustered. Most socio-economic information, such as population and traffic, is considered as variations in 2-dimensions [(X,Y) + attribute] only; whereas natural phenomena, such as temperature and pressure, vary with space and height [(X,Y,Z) + attribute]. When 'time' is added, it may be treated as merely an attribute to 2-dimensional or 3-dimensional space, for example, a date when a certain event occurs or a record is created; but this does not allow clustering in terms of time. An alternative common method is to model 'time' as a third dimension in addition to the 2-dimensional [(X,Y, T) + attribute] space. Therefore, some ST clustering methods have been developed from spatial clustering methods (Birant & Kut, 2007; Y. Hu et al., 2018; Lee et al., 2017; Tango et al., 2011). The addition of a time dimension to the 3-dimensional [(X,Y, T, Z) + attribute] space is still a challenging issue to model and to visualize. There is a need in many applications to integrate spatial and temporal information together for more detailed and accurate analyses. For example, in the study of human mobility, there is a need to identify at what time and where people cluster instead of just relying on census data or a generalized pattern of population distribution. This applies in the same way to crime patterns, traffic patterns etc. In the following sections, we will discuss the different categories of ST clustering.

2.3 Hypothesis testing-based methods

In the field of statistics, some existing fundamental research has been studied (Cressie & Wikle, 2015), including ST point pattern detection and analysis (Diggle, 1990, 2013). Hypothesis testing is used to determine the probability of a given hypothesis being true or not. The advantage of this method is it considers space and time information together. It is a new research direction that could allow some traditional spatial statistics to be extended for ST data analysis. For example, Di Martino and Sessa (2011) proposed an extended algorithm of fuzzy c-means to find circular clusters from ST data. This method could reduce the noise and outliers influencing clustering results. Detailed processes of some famous algorithms are described below.

2.3.1 Space–time interaction methods

A number of methods have been explored for detecting ST clustering. The core essence of a cluster is that objects should be close to each other in the space or time dimension. E. Knox and Bartlett (1964) proposed a test to quantify a space and time interaction of disease. Low-intensity disease detection by joining space and time analysis was conducted in Reference (G. Knox, 1963). Improvements to existing drawbacks were proposed by others (Kulldorff & Hjalmars, 1999). In this method, critical space distance α and time distance β should be manually defined first. Pairs of cases less than the critical space distance and time distance separately were regarded as near in space and time. The test statistics equation was:

$$K = \sum_{i=1}^{N} \sum_{j=1}^{i-1} d_{ij} t_{ij} \tag{2.1}$$

where K was the total number of paired cases smaller than the critical space and time distance, N was the total number of data. d_{ij} was space adjacency, if the distance between i and j was less than α, it was equal to 1, otherwise equal to 0. t_{ij} was time adjacency, if the distance between i and j was less than β, it was equal to 1, otherwise equal to 0. The Monte Carlo method was used for the significant test of K and a predefined number of runs was identified. The probability value of K being larger than the test statistic should belong to right hand tail of null distribution. The disadvantage of this method was critical space and time distances values may be assigned subjectively.

A modification was proposed by Mantel (1967) who multiplied the sum of time distances by the sum of spatial distances. The test statistic of Mantel's test was similar to Knox's test. It focused on the problem of selecting the critical distances of Knox's test. It is based on a simple cross-product term:

$$Z = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}^{s} d_{ij}^{t} \tag{2.2}$$

where d_{ij}^s is the distance between data i and j in space. d_{ij}^t is the distance between data i and j in time. Then, it is normalized:

$$M = \frac{1}{(N^2 - N - 1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\left(d_{ij}^s - \overline{d^s}\right)}{s_s} \frac{\left(d_{ij}^t - \overline{d^t}\right)}{s_t} \tag{2.3}$$

where M is the standardized Mantel statistic and N is the number of data. d_{ij}^s is the distance between data i and j in space. d_{ij}^t is the distance between data i and j in time. $\overline{d^s}$ is the average distance of all data in space. $\overline{d^t}$ is the average distance of all data in time. s_s and s_t are the standard deviations of data in space and time, respectively. This equation allowed for different units of space and time in the same framework, and multiple scale problems could be solved by limiting the range of correlation coefficient values into [–1,1].

2.3.2 Spatiotemporal k Nearest Neighbors Test

Jacquez (1996) proposed a spatiotemporal k nearest neighbors test to test space and time simultaneously. The statistic counted the number of k nearest neighbors in space and time dimension and evaluated under the null hypothesis of independent in two dimensions. Two test statistics were defined, which are D_k and ΔD_k. D_k is the count of case pairs of k nearest neighbors. It is large when space and time interact. ΔD_k is the count number of difference between consecutive k nearest neighbors. Some concepts are as follows:

N: Number of cases.

d_{ij}: Spatial measure, when $d_{ij} = 1$ case j is a k nearest neighbor of case i in space, otherwise equal to 0.

t_{ij}: Spatial measure, when $t_{ij} = 1$ case j is a k nearest neighbor of case i in time, otherwise equal to 0.

D_k: Is a cumulative test statistic, where $D_k = \sum_{i=1}^{N} \sum_{j=1}^{N} d_{ij}\, t_{ij}$.

ΔD_k: Is a k specific test statistic, where $\Delta D_k = D_k - D_{k-1}$.

D_k was not independent because it included a smaller k value of nearest neighbor. ΔD_k was independent because it only contained specific k nearest neighbors. The null hypothesis was that the distribution of events was independent from each other in space and time. Reference distribution was built by repeating many times to generate a random distribution for testing the statistics of probability values by comparing D_k and ΔD_k. However, the disadvantage of this method was that the k value could result in different test results

2.3.3 Scan Statistics

Scan statistics is a popular method and software (Kulldorff, 2018) can implement scan statistics for detecting clusters. Joseph Naus (Glaz et al., 2001) has been called the father of scan statistics as his method has helped to solve many research problems. The space scan statistic was developed from an original scan statistics method based on the scanning window process (Kulldorff, 1997). A circular scan window with different radii is used to find circular clusters of two-dimensional spatial data with a statistical significance test. An appropriate radius is important to avoid too large or too small clusters, otherwise the results could be meaningless and hard to interpret. Normally, the upper limit of the circle should not include more than 50 percent of all the dataset. Each point could be the center of a circle that contains different numbers of other points. Space and space–time scan statistics have many similar calculation processes.

Space–time scan statistics was extended from space scan statistics to detect clusters with the highest likelihood ratio by moving a cylinder as a scan window to scan ST data (Kulldorff, 2001; Kulldorff et al., 2005). Figure 2.3 shows the difference between the two methods. The left graph uses space scan statistics to detect clusters, the red center is the core point and the larger circle is the scan window for detection. The right graph uses space–time scan statistics to find clusters, it adopts a red cylinder as the scan window. Space–time scan statistics considers the time dimension and is an extension of space scan statistics in that a three-dimensional cylinder instead of a two-dimensional circle is used. The time interval between events is the height of cylinder.

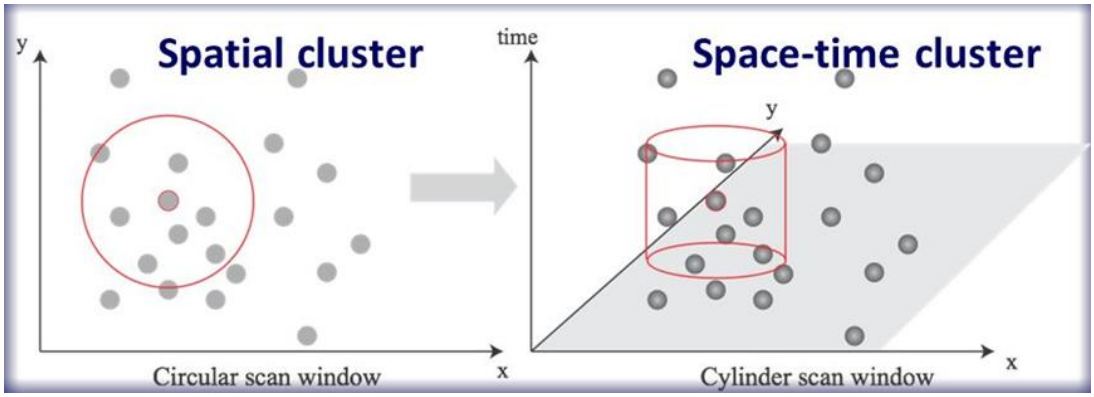


Figure 2. 3 Space and space–time scan window for detecting clusters (sources: Kulldorff (1997); Kulldorff et al. (2005)).

As with the space scan statistic, the null hypothesis is that the spatiotemporal distribution of events is random. The scan window of the cylinder was changed with different radii and height, looking for the maximum value of log likelihood ratio of all the circles as the cluster region. The formulation was:

$$S = log\left(\frac{n_z}{u_z}\right)^{n_z}\left(\frac{N-n_z}{N-u_z}\right)^{(N-n_z)} I\left(\frac{n_z}{u_z} > \frac{N-n_z}{N-u_z}\right) \tag{2.4}$$

where S was the log likelihood of cylinder, n_z and u_z were the observed and expected number of points, respectively, N was the total number of observed points, and I was the indicator function. If the left side was larger than right side, I was equal to 1, otherwise equal to 0. Many distribution functions could be used, one of which was the Poisson distribution. To obtain the simulated distribution for significance testing of clusters, Monte Carlo replications of data were used to obtain likelihood ratio statistics S. It was necessary to obtain p values by generating replications such as 999 or even higher to calculate the probability of a random appearance of an observed high-density cluster in a cylindrical window. The likely clusters could be based on the lowest p value, which was defined by the cylindrical window. However, similar to space scan, the disadvantage of this method was that it could not discover the arbitrary shape of ST data. To overcome this problem, flexible spatial scan statistic (Tango & Takahashi, 2005) and flexibly shaped

space–time scan statistic (Kunihiko Takahashi et al., 2008) were proposed in 2005 and 2008, respectively. FleXScan (K Takahashi et al., 2013) is the software that was developed to analyze spatial data by using flexible spatial scan statistics. Compare with spatial and space–time scan statistics that can only detect circular or cylinder clusters with variable size, these two methods have the ability to detect non-circular and non-cylinder clusters with high accuracy. For example, Tango and Takahashi (2005) proposed a flexible spatial scan statistics method that was illustrated using simulated disease maps in the Tokyo Metropolitan area. First, they divided the entire area into many small regions and the location of each region was the administrative population centroid. Next, the set of irregularly shaped windows were consisted K concentric circles and connected regions, where K is a pre-specified maximum length of cluster. The idea was also used in the flexible space–time scan statistic. However, both of these were fitted to a small cluster size. Neill (2006) gave a very comprehensive account of spatial and ST clustering methods, especially in the area of scan statistics methods and Bayesian clustering methods. They proposed a statistical framework for detecting clusters in detail. The results of case studies show it has good performance compared to previous studies. However, they are still subject to the limitations of statistical methods.

2.4 Partitional Clustering Methods

In the previous section, clustering of hypothesis testing-based methods was developed based on mathematical theory of probability and statistics. In this section, partitional clustering methods are introduced. These methods mainly focus on identifying whether data belong to a cluster or noise by using different distance functions. They have a clear grouping process to form a cluster by determining the similarity of data. Some well-known methods are described as follows:

2.4.1 DBSCAN

DBSCAN is a very popular method, especially in the data mining community (Han et al., 2011; Miller & Han, 2009). It has been extended for many different types of data. The biggest advantages

of this method is that it can find clusters with arbitrary shape and noise points (Ester et al., 1996). The key idea is that each cluster should include at least a minimum number of points with a fixed radius. Similar to KDE, DBSCAN can also be extended for spatiotemporal data. ST-DBSCAN (Birant & Kut, 2007; M. Wang et al., 2006) was proposed to cluster spatiotemporal data. M. Wang et al. (2006) added another radius r_t which is the temporal neighborhood radius. The core points should satisfy directly the density reachable in both spatial radius r_s and temporal radius r_t.

To define an appropriate spatial and temporal radius, k-dist graph was used to decide values. Generally speaking, cluster data should be clearly separated from noise data. To do this, the distance of each point to its k nearest neighbor, called the k value, was calculated. As depicted in Figure 2.4, the left graph shows the distribution of point sample, clearly indicating three similar density clusters surrounded by noise points. The right graph was drawn based on a descending order of k values. The smooth red line on the right part of the graph highlights cluster points that have a low k value, but the left part of the red line indicates noise points that have high values. An appropriate threshold could be selected from the graph with an obvious and abrupt change from high value of small number of points to low value of large number of points.

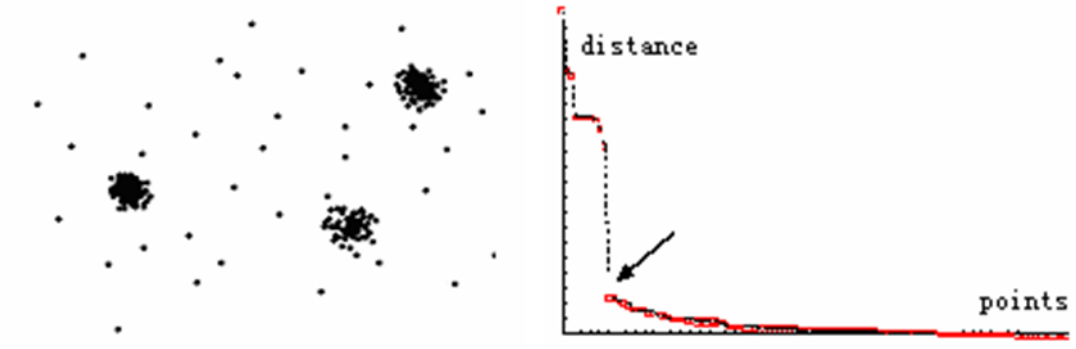


Figure 2. 4 A point set with its sorted 4-dist graph (source: M. Wang et al. (2006))

Another method was called ST-GRID. The core idea was that a three-dimensional grid covers the entire dataset followed by merging the dense neighboring cells. First, the above k-dist graph could be used to define the border length of the grid and put all the data into a multi-dimension grid. Second, the number of points in each cell was counted. Those equal or larger than $k+1$ were

merged with neighbor cells as a cluster. The process was repeated until no additional cells could be merged.

Compared with the above method, more detailed data such as non-spatial data should be considered when extending DBSCAN (Birant & Kut, 2006, 2007). A new method called ST-DBSCAN was proposed for discovering clusters based on three attributes; non-spatial, spatial, and temporal attributes of data. Basic concepts were the same as conventional DBSCAN except for three modifications.

When DBSCAN only considers one distance parameter to find similar data, ST-DBSCAN used two distance parameters for two-dimensional data. One distance measured two points distance in spatial scale. Another distance measured non-spatial attributes. Euclidean distance was adopted to calculate the two distances.

$$Eps = \sqrt{(x1 - x2)^2 + (y1 - y2)^2} \tag{2.5}$$

where x and y represented spatial information. DBSCAN algorithm' result could be affected by selecting a different radius. If the dataset included different densities of clusters, a single radius could not clearly identify each cluster. To solve the problem, they proposed a concept called the density factor. Each cluster has their own density factor. To calculate it, three concepts of distances are introduced, which are density_distance_max, density_distance_min and density_distance. Density_distance_max was the maximum distance between object p and its neighbor objects within the radius Eps. Density_distance_min was the minimum distance of each cluster. The density_distance of object p was defined as density_distance_max (p)/density_distance_min (p). The density_factor was defined as follows.

$$Density\_factor(C) = {}^{1}\Big/\left[\frac{\sum density\_distance(p)}{|C|}\right] \tag{2.6}$$

The density_factor C denoted the degree of each cluster. If the points of a cluster were close to

each other, density_diatance_min would decrease, the density_distance would be quite large, and the density_factor would be close to 0. Otherwise, if points were a little further away from each other, the density_distance would be quite small and the density_factory would be close to 1.

For non-spatial values of objects, this added value could change the average value of existing points when clustering. To solve the problem, ST-DBCSAN compared the average value of a cluster with every other point. If the absolute difference between the average value and object value was larger than a threshold, that point should not to be contained in the cluster.

2.4.2 Kernel Density Estimation

Kernel density estimation (KDE) (Scott, 2015; Silverman, 1986) is a nonparametric density estimation method widely used for detecting clusters from spatial data to discover high-density significant geographic events. Gaussian function is an efficient and popular choice for kernel density estimation. The KDE equation can be extended as follows:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2} \tag{2.7}$$

where n was the number of sample data, h meant the bandwidth parameter, and K was the kernel density functions. Many kernel functions had been defined for different situations. An appropriate bandwidth could lead to a good density result. The function of Scott's rule of thumb was used to calculate bandwidth with the equation as follows:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5} \tag{2.8}$$

where $\hat{\sigma}$ was the standard deviation of sample data, and n meant the number of sample dataset. This rule of thumb was very easy to compute and could be accepted as an accurate estimator. There

are mainly two ways to extend KDE for spatiotemporal data by adding a time dimension (Table 2.1).

Table 2. 1 Comparison of different extension methods

Authors	Methods
Brunsdon et al. (2007) Nakaya and Yano (2010) Wei et al. (2018)	Temporal attribute is regarded as another dimension, calculate space and time kernel density estimations separately.
Lee et al. (2017)	Setting a threshold to filter inappropriate space and time distances. Standardization of space and time data for integrating them with same kernel function.

Conventional KDE should be extended by adjusting the parameters for spatiotemporal data. Brunsdon et al. (2007) extended the two-dimensional KDE into three-dimensional for space and time data analysis. It helped to visualize and understand the trend of spatiotemporal data. The three-dimensional spatiotemporal KDE formula was:

$$\hat{f}(x,y,z) = \frac{1}{nh_s^2h_t}\sum_{i=1}^{n} k_s\left(\frac{x-x_i}{h_s},\frac{y-y_i}{h_s}\right)k_t\left(\frac{t-t_i}{h_t}\right) = \frac{1}{nh_s^2h_t}\sum_{i=1}^{n} k_s(u_s)\,k_t(u_t) \tag{2.9}$$

where k_t was the kernel function for time, h_t was the bandwidth parameter of time kernel. Spatial and temporal information were treated separately, each of which had its own bandwidths and kernel functions. Nakaya and Yano (2010) adopted this method for visualizing high-density crime events during a specific time interval in Kyoto. A threshold was set to filter data beyond a defined range. For most data, the longer space/time distance between two datasets, the lower possibility of their correlation. For example, if the time distance of two adjacent data was larger than a threshold, there was no need to calculate kernel density. The advantage of this method was no requirement to define a density function of time, but time was regarded as a constant. The formula was:

$$\hat{f}(x,y,z) = \frac{1}{nh_s^2}\sum_{i=1}^{n} k_s\left(\frac{x-x_i}{h_s},\frac{y-y_i}{h_s}\right), u_t < h_t = \frac{1}{nh_s^2}\sum_{i=1}^{n} k_s(u_s), u_t < h_t \quad (2.10)$$

In this formula, only kernel density of space needs to be calculated. However, it is difficult to define an appropriate method for filtering time. In order to directly integrated space and time data, the process of standardization should be conducted before density estimation with the following equations:

$$s = \frac{s' - \bar{s}}{h_s} \quad (2.11)$$

and,

$$t = \frac{t' - \bar{t}}{h_t} \quad (2.12)$$

where s', t' were spatial and temporal raw data, $\bar{s}$, $\bar{t}$ could be referenced values for standardizing raw spatial and temporal data and h_s,h_t were their kernel bandwidths. The advantage of standardization of raw spatial and temporal data was to remove the different measurement units of spatial and temporal data. The results of standardization of spatial and temporal data was that they have similar ranges for easy integration. The calculation of kernel density estimation was

$$\hat{f}(x,y,z) = \frac{1}{nh_s^2 h_t}\sum_{i=1}^{n} k_{st}(u_{st}) = \frac{1}{nh_s^2 h_t}\sum_{i=1}^{n} k_s\left(\frac{x-x_i}{h_s},\frac{y-y_i}{h_s},\frac{t-t_i}{h_t}\right) \quad (2.13)$$

$$u_{st} = \sqrt{\left(\frac{x-x_i}{h_s}\right)^2 + \left(\frac{y-y_i}{h_s}\right)^2 + \left(\frac{t-t_i}{h_t}\right)^2} \quad (2.14)$$

However, it is noted that bandwidth selection was a critical problem that will affect cluster results. The unit of time was another problem because different units lead to different density of clusters.

2.4.3 Windowed Nearest Neighbor Method

Based on the idea of spatiotemporal k nearest neighbors test, windowed nearest neighbor method for mining spatiotemporal clusters was proposed several years ago (Pei et al., 2010). Spatiotemporal point data could be represented by ST_p, each point indicated by $ST_p(s_i, t_i)$, and its neighbor could be defined as:

$$ST_p = \{ST_p(s_i, t_i), ST_p(s_{i+1}, t_{i+1}), ST_p(s_{i+2}, t_{i+2}), \dots\dots, ST_p(s_{i+n}, t_{i+n})\} \tag{2.15}$$

For k nearest neighbors, the time interval of consecutive two points should be smaller than a threshold, $|T_{i+1} - T_i| \leq \Delta T$. The distances $D(ST_p)$ from a given point to the rests are gradually increasing with time satisfied as:

$$\begin{aligned} D\left(ST_p(s_i, t_i), ST_p(s_{i+1}, t_{i+1})\right) \leq D\left(ST_p(s_i, t_i), ST_p(s_{i+2}, t_{i+2})\right) \leq \cdots \\ \leq D\left(ST_p(s_i, t_i), ST_p(s_{i+k}, t_{i+k})\right) \end{aligned} \tag{2.16}$$

Similar to space–time scan statistics, each event could be regarded as a center of cylinder with a spatial radius and temporal height. A cylinder as a window includes spatiotemporal neighbors of a given event. A core event's neighbor should contain a minimum number of other points. The first step is to distinguish between a cluster of events and noise; second is to connect the cylinder into cluster events. Figure 2.5 shows the spatiotemporal density connectivity of events from a horizontal perspective to form the cluster.

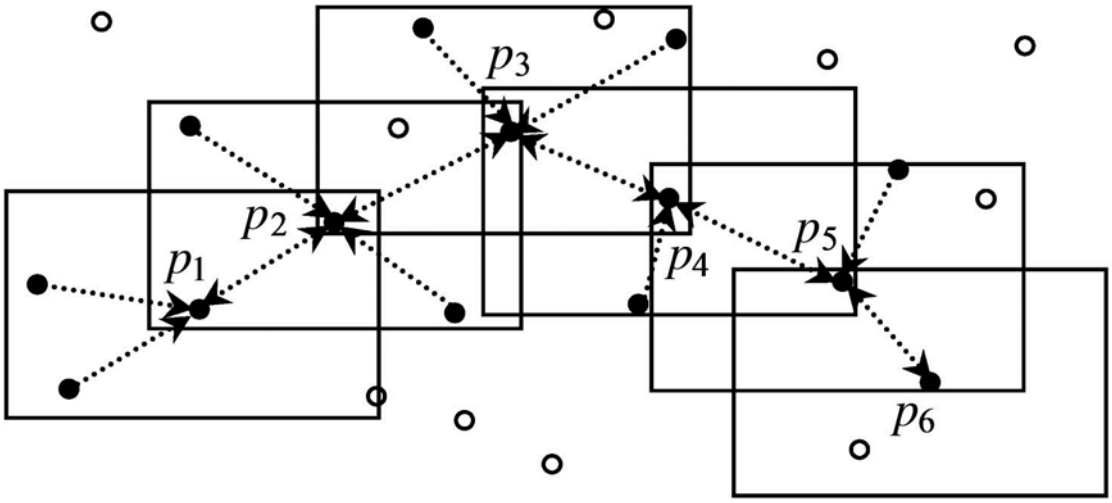


Figure 2. 5 Spatiotemporal density connectivity (source: Pei et al. (2010))

In their method, an ST Poisson point process was used to construct probability density function with the equation:

$$P(k) = \frac{\lambda^{kV}}{k!} e^{-\lambda V} \tag{2.17}$$

where k was the number of events in the volume of V, λ was the constant. The density of cylinder D can be calculated by:

$$D = \frac{k}{\pi \Delta S_t \Delta T} \tag{2.18}$$

where k was the number of events, and ΔT was the temporal interval constant. ΔS_t could be regarded as a threshold calculated by an expectation maximization (EM) algorithm (Dempster et al., 1977). A detailed process of the EM algorithm can be found in Byers and Raftery (1998). After the density connected events were divided into cluster events and noise features, they were linked by the cylinder for connecting events into clusters.

2.5 Applications

ST data clustering methods are widely used in many research areas, which we have divided into the following categories.

Crime analysis: Criminal events usually repeatedly occur under the same situation and a similar time. Tracing the changes of a crime path is meaningful. Nakaya and Yano (2010) explored the possibility of tracing crime events with three-dimensional attributes in a space–time cube relative to kernel density estimation and scan statistics to get the clustering of crime events and visualizing the crime events patterns. Y. Hu et al. (2018) proposed a new modification of an existing method to increase the predictive accuracy of crime hotspots. They refined spatiotemporal kernel density estimation by generalized product kernels and adopted a data driven bandwidth selection to decide bandwidth. Residential burglaries data of Baton Rouge was used to predict crime hotspots.

Events detection: Many events could be detected using clustering methods, such as helicopter crash accidents from social media data (T. Cheng & Wicks, 2014). By using space–time scan statistics, a ST significant cluster of London helicopter crash locations were found. Many other events like football games and train and flight delays could also be detected. Clustering earthquake events could help to understand trends and mechanisms (Chen et al., 1999; Ripepe et al., 2000). Many small earthquakes can happen before or after a strong earthquake. By using ST clustering methods, clusters of earthquakes can be identified in space and time. ST kernel density estimation can be used for predicting ambulance demand. It is difficult to predict ambulance demand accurately from large-scale datasets of past events. Zhou and Matteson (2015) proposed a model of spatiotemporal kernel density predictive method to explore ambulance demand precisely. KDE is also widely used in creating a density map of road accidents to identify its distribution pattern (Anderson, 2009). This could help to predict and reduce the number of incidents in the future.

Mobility: Human mobility data such as phone call data could reflect urban growth in space and time. It would provide information for authorities to plan and manage cities in a smart way. It helps

planners to understand where and when different groups of people interact in urban space. Jiang et al. (2012) discovered the clusters of human mobility pattern by kernel density estimation and integrating various spatial and temporal data to predict human daily routines. Krisp et al. (2013) proposed directed kernel density estimation to recognize movement and direction of crowds and was effective in visualizing the movement of crowds.

Disease analysis: ST clustering methods could be applied in analyzing disease dispersion and trends. Visualizing space–time clusters of dengue fever pattern in Cali using extension of kernel density estimation method has been applied (Delmelle et al., 2013; Delmelle et al., 2014). The occurrence and spread of disease has a strong regular pattern in certain regions. Analyzing the former spread of disease to predict the future spread direction is meaningful for governments and hospitals to control diseases. Gomide et al. (2011) analyzed not only the location and time the disease was contracted, but also the reaction of the population when facing the disease. They used the ST-DBSCAN clustering method to explore the ST distribution characteristics of disease incidents to group nearby cities that have similar incident rates. A linear regression model was built to predict the number of diseases using the proportion of user experiences. Napier et al. (2018) proposed a novel Bayesian model to identify the cluster of similar temporal disease trends rather than disease estimation and prediction. Adin et al. (2018) proposed a two-stage approach to estimate disease risk maps. Compared with traditional methods, their method has the ability to overcome the problem of local discontinuities in the spatial pattern that cannot be modeled. It has a good performance of spatiotemporal smoothing for estimating risks of disease mapping.

2.6 Conclusion

ST data clustering analysis is a hot topic and has already been studied extensively (Shekhar et al., 2015). ST data types can be classified into three categories, namely point, line, and polygon. In this chapter, only point pattern is considered and existing clustering methods are divided into two parts, one is hypothesis testing based, and another is partitional clustering methods. ST data is more complicated than other types of data because of the additional dimension of time from two-

dimensional spatial analysis. Some popular and representative methods are introduced in previous sections. However, simply regarding time as an extended dimension may ignore some important patterns that are hard to be detected. New methods should consider integrating time and other attributes together.

Clustering is an important step to detect patterns from a large amount of data. It can be used in many application domains, including transportation, social media, and urban development. It focuses on finding hotspots from raw data. These hotspots are the foundation for pattern understanding. Adjusting different parameters of the clustering method for different data types is needed to get an optimum result. An appropriate clustering method can help discover potential and useful information from a large volume of data. Apart from investigating new algorithms, related research problems have been developed, such as the computational issues of ST data (Vatsavai et al., 2012). As mentioned before, even though extended algorithms could be used to detect clusters, these are mere geometrical considerations. There is a need to predefine thresholds such as radius, distance, and density based on the rules or knowledge from specific themes. As such, new research trends and methods need to be developed.

ST data analysis has attracted much research attention and a lot of methods have been developed (Shekhar et al., 2008). However, there are still some issues and challenges to be solved. Several challenge issues are described as follows:

1. Multiple scales clustering of ST data is an important research topic. Clustering results could be different with both changing map scales and data scale of nominal, ordinal, interval, or ratio value (i.e., with increasing attribute information). The problem of multiple scales is related to different shapes, sizes, and densities of event distribution. Identify the multiple shapes and densities features is difficult.

2. Cluster number is difficult to estimate correctly without rich experiences. It is hard to determine the cluster number from the dataset by users.

3. It is difficult to determine the parameters of clustering method. Several clustering methods are needed set some parameters when clustering. However, it is difficult for users to set the parameters.

4. Different types of ST data analysis should be considered to develop diverse clustering methods. In many existing studies, most algorithms are focused on point features or events. However, spatiotemporal data from GPS and other positioning equipment can record locational information in a linear dimension, thus demanding new methods for line clustering. The same applies for outliers' detection (T. Cheng & Li, 2004, 2006) and classification algorithms that have not been investigated thoroughly yet.

5. Different patterns could result from using different spatial or temporal attributes. It is difficult to detect the best pattern based on one algorithm. Generally speaking, raw data could contain many different kinds of pattern. For efficient mining of potential patterns, new algorithms for evaluating the accuracy or reliability of various patterns should be investigated in the future.

Chapter 3 Voronoi construction based on an integrated clustering method for region partition

Chapter 2 systematically review the current clustering methods and identify the limitations of the existing methods. In this chapter, Voronoi construction based on an integrated clustering method was proposed. The integrated clustering method can detect the multiple densities and shapes feature. Two improvements are made regarding the limitations of existing clustering methods. First, the number of clusters can be determined automatically, which is an improvement from current clustering methods that have difficulty determining the number of clusters. Second, the integrated method is easy to implement and requires only one parameter which is needed to be set by user. The parameter is used to describe distance in the clustering method. It is therefore more efficient and reliable than existing methods that require more parameters, which are normally determined by the users themselves. These two improvements allow the integrated clustering method to detect multiple densities and shapes of clusters. The Voronoi diagram is used to partition region. This method is suitable for different kinds of data sources.

3.1 Introduction

Clustering is the process of dividing objects into multiple groups. Each group is called a cluster, and each cluster includes several similar objects. With the development of big data, clustering has emerged as a powerful data mining method to detect hot spots that include valuable thematic information.

The detection of clusters for datasets with different shapes, densities, and sizes, even including noise, remains an essential and open issue. Various type of techniques, such as density methods and

hierarchical methods, and even their improved versions, are commonly used for accurate detection of clusters, but several challenges remain. First, it is difficult to predefine the number of clusters, such as the “k” in k-means clustering. Second, many methods require more than one parameter, which is difficult to determine appropriately for many cases. The choice of parameters can significantly affect the cluster results. For example, two parameters of DBSCAN must be set manually. At the same time, if an appropriate start point is not selected, some points could be regarded as noise points, for example if the start point is selected from the low-density points. Rodriguez and Laio (2014) proposed a fast clustering method. The core idea is that the centers of clusters are surrounded by points with a certain relative density but far away from others that have high density. The centers of the clusters are based on the density functions and density radius. Theoretically, this method does not require the parameters to be predefined. The centers of clusters can be calculated based on the product of density and distance. However, this method cannot automatically select the cluster centers, so they must be decided by observation.

In this chapter, Voronoi construction based on an integrated method is proposed for region partition. The method is divided into two parts, an integrated clustering method and Voronoi diagram construction. Two improvements of the clustering method are made regarding to the limitation of the existing clustering methods. First, cluster number can be determined automatically, which is a step further to the current clustering methods with the difficulties of determining cluster number. Second, the integrated clustering method is easy to implement and only one parameter is to be provided. It is therefore more efficient and reliable than the existing methods where more than the case where more parameters to be provided, normally determined by users themselves.

With the above mentioned two improvements, multiple densities and shapes of clusters can be detected by the clustering method. The integrated clustering method can be applied to various datasets, and it performs better than the existing clustering methods, especially for the datasets with multiple densities and different shapes. Both significant and weak cluster patterns can be discovered from the clustering results by using this method. The center points of clusters are used as seed to construct the Voronoi diagram. Figure 3. 1 shows the logical flow of the method.

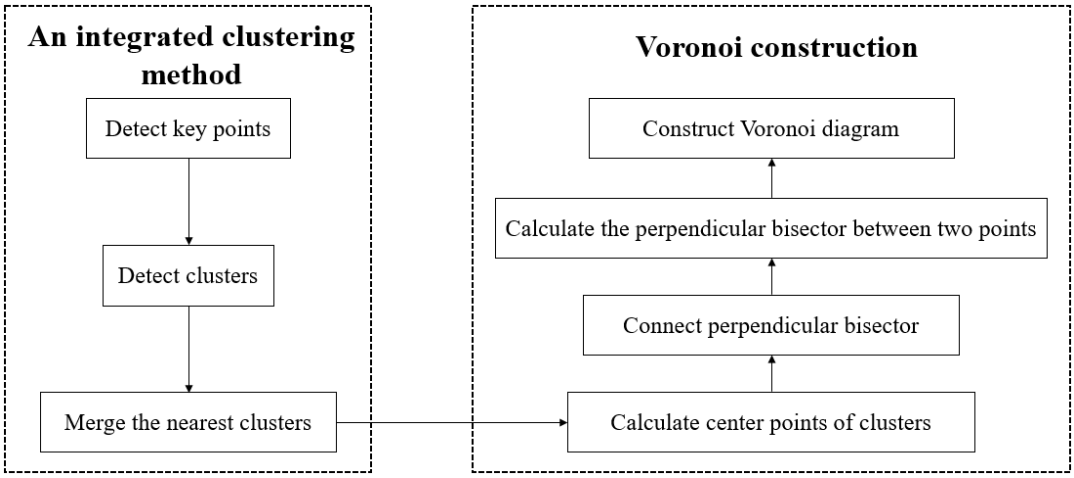


Figure 3. 1 Logical flow of Voronoi construction based on the integrated clustering method

3.2 Integrated clustering method

3.2.1 Logical flow of integrated method

Considering the limitations of the existing methods, we proposed an integrated clustering method to determine clusters from datasets with multiple densities and shapes. Logically, the method includes three main steps: detection of key points, detection of clusters, and merging of the nearest clusters. First, points are calculated with high kernel values and selected to be key points for a preliminary determination of the number of clusters. Second, based on the number of cluster points, clusters are detected with the identified hierarchical method. Third, the clusters that are near each other are merged to form a larger cluster. Figure 3. 2 shows the logical flow of this method. The details of each step are described in the following sections of this chapter.

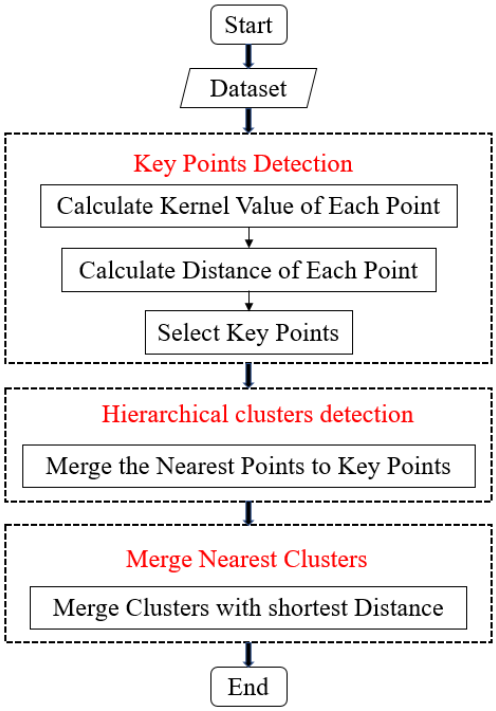


Figure 3. 2 Logical flow of the proposed integrated clustering method

3.2.2 Definition of relevant concepts

To illustrate this method, several concepts relevant to points are defined here and are used in various steps. This section gives the basic relevant definitions of the concepts. Related information is introduced in later sections. Figure 3. 3 illustrates a cluster in the process to help to explain the following concepts.

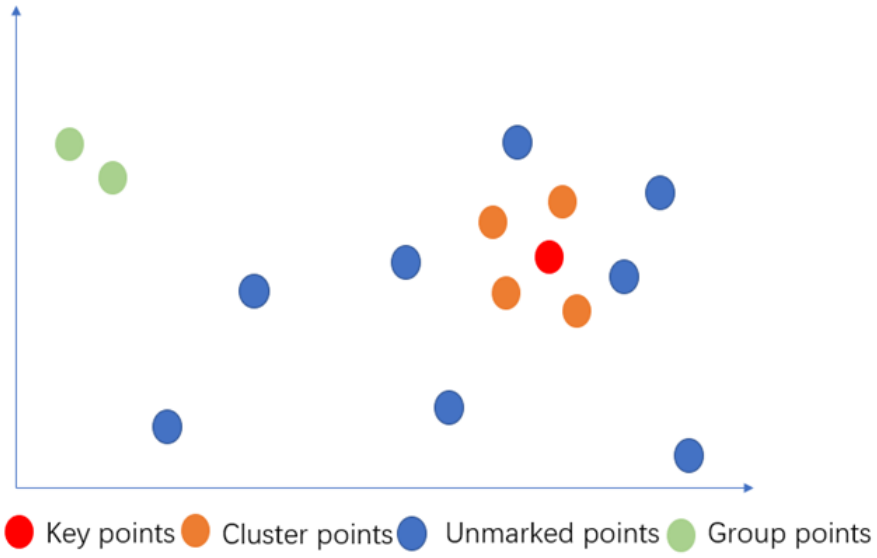


Figure 3. 3 Categories of points

1) Key points

Key points are points that have been marked as a member of one cluster firstly. In this method, several key points are first selected according to distance and local density, and the number of key points is equal to the number of clusters. When a hierarchal method is used to detect clusters, if an unmarked point is near the cluster point, it becomes a part of that cluster. In Figure 3. 3, key points are shown in red.

2) Cluster points

When a hierarchal method is used to detect clusters, some points are near the key points. These points are called cluster points. In Figure 3. 3, cluster points are shown in orange.

3) Group points

When a hierarchal method is used to detect clusters, some points are near each other but far from key points or cluster points. These points can be grouped together and called group points. In Figure 3. 3, group points are shown in green.

4) Marked points and unmarked points

Marked points include key points and cluster points, and unmarked points are those not categorized

as key points, cluster points, or group points. In Figure 3. 3, unmarked points are shown in blue.

3.2.3 Detection of key points

The core idea of detecting key points is to determine each cluster's high-density point, that is, the point surrounded by several points within a short distance. To determine the points, each point is computed for its two attributes (the local density and the distance from the point of high density). Two main methods are used to compute the local density. One is defined as follows:

$$D_i = \sum_j \lambda(d_{ij} - d_s) \tag{3.1}$$

where D_i is the local density of point i, $\lambda(d_{ij} - d_s) = \lambda(x) = 1$ if $x < 0$ or $x = 0$; otherwise $\lambda(x) = 0$, d_s is the specified cutoff distance. The local density is the number of points with a shorter distance than d_s to point i. The other method is to use the kernel function to calculate the local density. It is defined as follows:

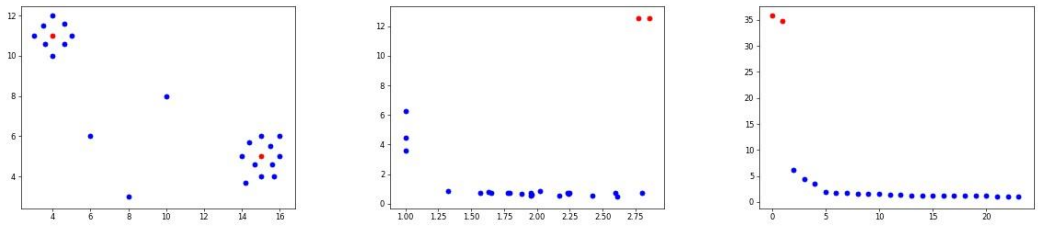
$$D_i = \sum_j e^{-\left(\frac{d_{ij}}{d_s}\right)^2} \tag{3.2}$$

Compared with the cutoff method, which generates random values, the kernel function obtains continuous values and each point can be given different values. There is no doubt that the method of choosing an appropriate d_s is important. In this chapter, we refer to the work (Rodriguez & Laio, 2014) and select the 2% distance as the d_s.

The next step is to compute the distance from high-density points to low-density points. First, the points are sorted in descending order based on the local density. Second, the shortest distance from points with a larger value of local density to point i is calculated as the distance β_i. This process begins from the point with the second-highest local density because the first point has the highest

local density. The point with the greatest distance is set as the first point.

From now on, each point has two attributes: a) local density D_i and b) distance β_i. To determine the centers of the clusters, the two attributes must be plotted in two-dimensional coordinates. In Figure 3. 4(b), the x-axis is the local density and y-axis is the distance.



(a) (b) (c)

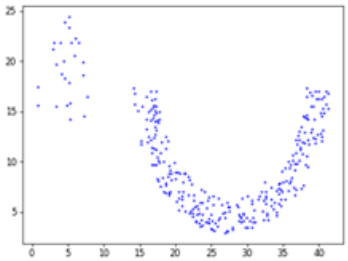
Figure 3. 4 (a) Distribution of dataset; (b) Distribution of local density and distance; and (c) Product result of local density and distance

Figure 3. 4 (a) is the sample dataset, Figure 3. 4 (b) is the plot of the local density and distance for each point. Point with larger values of local density and a long distance to other high-density points are shown in red. Specifically, the two red points in the centers of the clusters have the largest local density and distance, and stand in obviously distinct positions from the blue points in Figure 3. 4(b) and Figure 3. 4(c). Even though the points around the center points also have high local density, they have short distances. The calculated products of local density and distance for the rest of the points are clearly shown in Figure 3. 4(b) and Figure 3. 4(c). After plotting all points, the red points are recognized as the centers of the clusters.

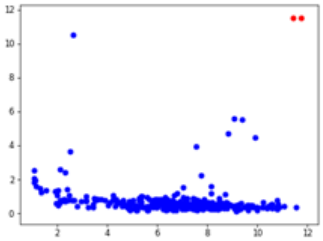
This method performs well with clusters with the same density and a spherical shape. However, the sample data set includes two clusters with different densities and a non-spherical shape, so it is difficult to locate their center. The local density must be calculated for each point and the results sorted to obtain the distance; the centers of high-density clusters would rank first, and those of low-density clusters would rank last. It is difficult to choose the correct number of center points. If a cluster has a non-spherical shape, several center points may be detected. For example, Figure 3. 5

shows the process of key point detection from a dataset with multiple shapes and densities. Figure 3. 5(a) shows an example of dataset distribution. It clearly shows that there are two clusters to be detected, and the two clusters have different numbers of points and densities. One cluster has a small number of points and a low density, and the other has a large number of points and a high density.

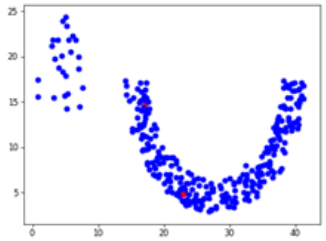
One obvious characteristic is that the right cluster has a U shape and includes more high-density points than the left cluster. Key points are detected with the method introduced above. Figure 3. 5(b) and Figure 3. 5(c) show the results of key point detection. Figure 3. 5(b) shows that the two points at the top right corner are the core points. Even though the number of key points equal to number of cluster, both of them are in the same cluster in Figure 3. 5(c). Nevertheless, the correct number of clusters can be detected from these results. In Figure 3. 5(d), three points are regarded as key points. In Figure 3. 5(e), the added key point belongs to a cluster that cannot be detected above. The comparison suggests that it is better to select as many key points as possible. Because the number of core points could exceed the number of clusters, merging the nearest clusters is the next step.



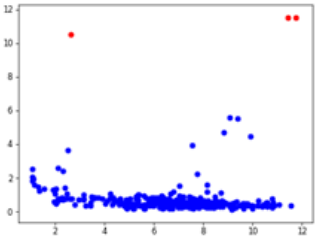
(a) Dataset distribution



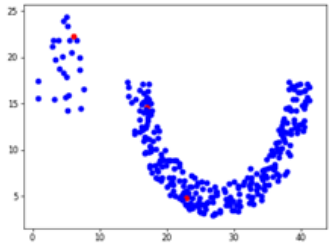
(b) Two points are selected



(c) The location of selected two points



(d) Three points are selected



(e) The location of selected three points

Figure 3. 5 Core point detection of dataset with multiple shape and densities

3.2.4 Cluster detection

After several key points are selected, an agglomerative hierarchical method is adopted to form clusters. This is a bottom-up process that merges the points to form clusters from small to large. The number of key points indicates the number of clusters that should be formed. Three phases are included in this step. First, the distance of all pairs of two points is calculated. Second, the distances are sorted from small to large. Third, the points are merged to form clusters according to distance. Table 3.1 is the pseudo-code of the developed clustering method, and Figure 3. 6 demonstrates the method's logical flow.

Table 3. 1 Pseudo code of developed clustering method

Algorithm: Cluster detection according to cluster points

Input: Dataset $P = \{p_1, p_2, p_3 \dots p_i\}$

Four kinds of point categories include key points, cluster points, group points and unmarked points $M = \{m_1, m_2, m_3, \dots m_i\}$

Key points $CP = \{cp_1, cp_2, cp_3 \dots cp_i\}$

Output: All points are cluster points

1: Calculate distances of each pair of points

2: Sort distances of pair of points by ascending order $D = \{d_{12}, d_{13}, d_{14} \dots d_{ij}, (i \neq j)\}$,

3: pairs of point $P = \{p_1p_2, p_1p_3, p_1p_4 \dots p_ip_j, (i \neq j)\}$

4: For i to D.length

5: If point $p_i \leftarrow$ key point or cluster point

6: If point $p_j \leftarrow$ unmarked point

7: p_j become same cluster point of p_i, $m_j = m_i$

8: else if point $p_j \leftarrow$ group point

9: p_j become same cluster point of p_i, $m_j = m_i$, rest points in the group

10: become same cluster of p_j, $m_g = m_i$

11: Else if point $p_j \leftarrow$key point or cluster point

12: If point $p_i \leftarrow$ unmarked point

13: p_i become same cluster point of p_j, $m_i = m_j$

14: else if point $p_i \leftarrow$ group point

15: p_i become same cluster point of p_j, $m_i = m_j$, rest points in the group

16: become same cluster of p_i, $m_g = m_j$

17: Else if point $p_i \leftarrow$ group point

18: If point $p_j \leftarrow$ unmarked point

19: p_j become same cluster point of p_i, $m_j = m_i$

20: Else if point $p_j \leftarrow$ group point

21: p_j become same group point of p_i, $m_j = m_i$, rest points in the group

22: become same cluster of p_j, $m_g = m_i$

23: Else if point $p_j \leftarrow$ group point

24: If point $p_i \leftarrow$ unmarked point

25: p_i become same cluster point of p_j, $m_i = m_j$

26: Else if point $p_i \leftarrow$ group point

27: p_j become same group point of p_i, $m_j = m_i$, rest points in the group

28: become same cluster of p_j, $m_g = m_i$

29: Else if point $p_i \leftarrow$ unmarked point and $p_j \leftarrow$ unmarked point

30: p_i and p_j become same group points, $m_i = m_j$

31: If all the points are cluster points

32: Cluster detection finished

33: Else

34: Next distance d_{ij}

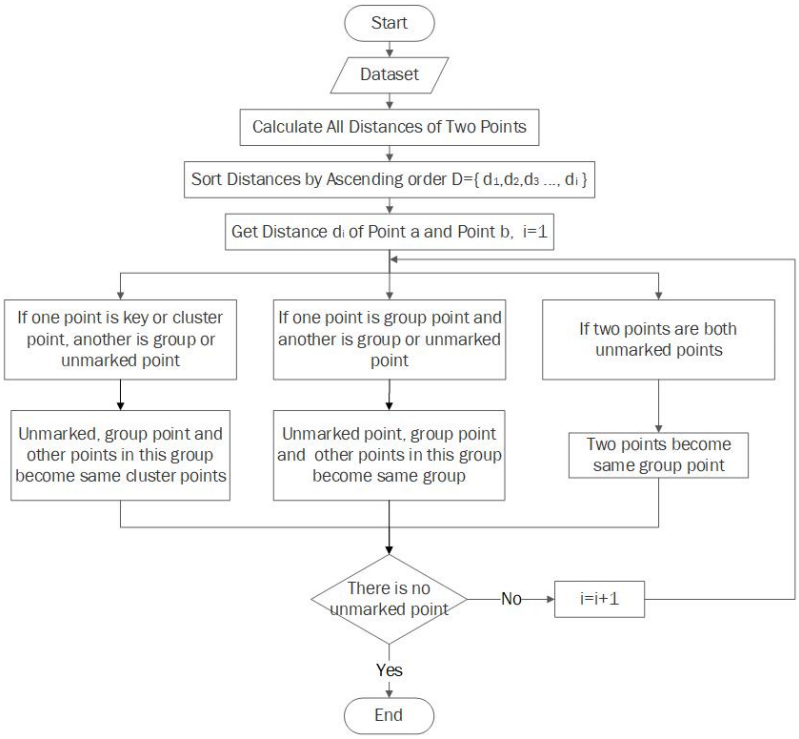
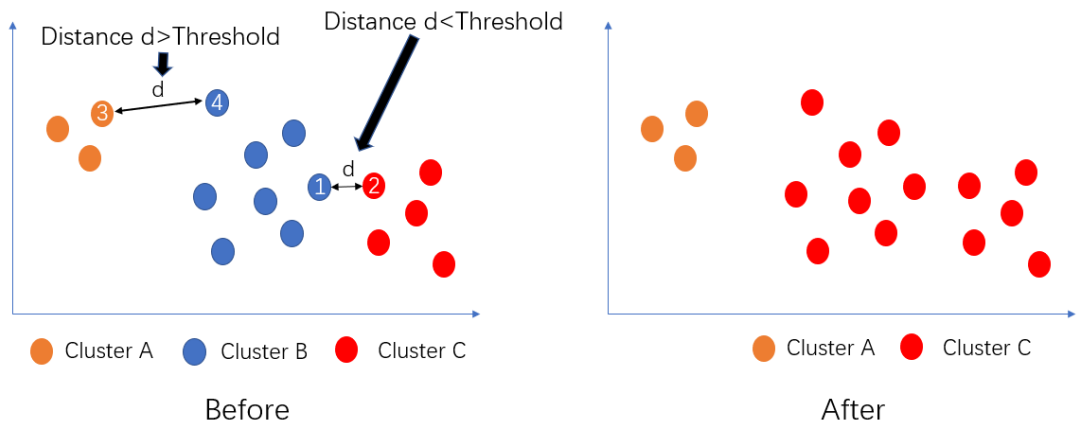


Figure 3. 6 Logical flow of cluster detection

For the first phase, an $n \times n$ table is built to record the distances of each pair of points when there are n points. The next phase sorts the points by distance in ascending order in terms of the former table. The third phase groups the two points to form clusters beginning from the shortest distance. Three situations can occur when grouping two points. First, when one point is a key or cluster point, the other is a group point or unmarked point that later becomes a cluster point. The other points in the same group become group points. Second, when one point is a group point, the other is a group or unmarked point that later becomes the same group point. Third, when both points are unmarked cluster points, they do not belong to any cluster. The two points are later marked in the same group, but they still do not belong to any cluster. This process begins from the first distance until no unmarked points remain.

3.2.5 Merging neighbor clusters based on threshold

After the clusters have all been detected by the hierarchical method based on the start point, the dataset has been divided into multiple clusters. However, some clusters are so near each other that they should be regarded as one. The final step in this part is that the nearest clusters are merged to form a larger cluster. A threshold for distance is set by the user to judge whether two clusters should be merged. If the shortest distance between two clusters exceeds the threshold, the clusters should not be merged; if the distance is below the threshold, the two clusters should be regarded as one cluster and merged. This merging process is repeated until no clusters need to be merged. Figure 3. 7 shows an example of this process.



(a) Clusters distribution before merge (b) Clusters distribution after merge

Figure 3. 7 Process of merging clusters

Figure 3. 7(a) shows three clusters in different colors detected with the hierarchical method. The nearest two points of two clusters are calculated and compared with the threshold value, which is set by the user, to decide whether the two clusters should be merged. Cluster A is near cluster B, and cluster B is near cluster C. The distance between points 1 and 2 is the shortest distance between clusters B and C. The distance between 1 and 2 falls below the threshold, so the clusters are merged. The distance between points 3 and 4 is the shortest distance between clusters A and B, but it exceeds the threshold, so these clusters should not be merged. Figure 3. 7(b) shows the resulting dataset, in which cluster B and cluster C become cluster C; therefore, two clusters are detected.

3.2.6 Distance measurement

Distance is a key concept in clustering methods and is normally used to measure the similarity of two objects in the clustering process. Similar objects in the same cluster are closer to each other than objects in other clusters, and they have a shorter distance when within a cluster than when between clusters. Objects are grouped in the same clusters due to the short distance between them. The improved cluster method proposed in this thesis has no limitations regarding the distance measurement that should be used. The distance measurement model for the proposed improved clustering method should be determined based on the specific case concerned. Table 3.2 summarizes the commonly used distance measures, which form the basis for the use of the proposed clustering method.

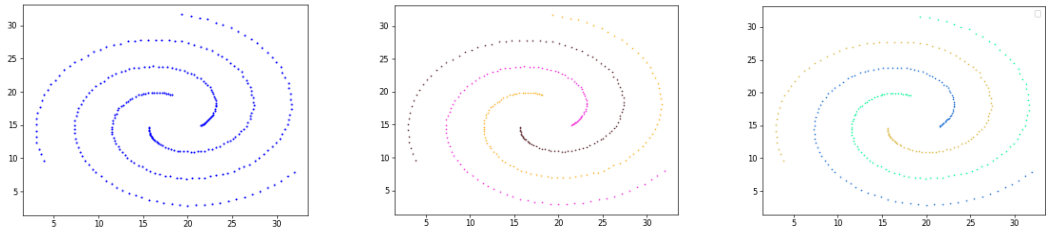
Table 3. 2 Distance measures

Name	Formula	Suitability
Euclidean Distance	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	Distance of straight line between two points
Manhattan Distance	$\sum_{i=1}^{n}\lvert x_i - y_i\rvert$	Distance between city blocks
Mahalanobis Distance	$\sqrt{(x-y)^T S^{-1}(x-y)^T}$	Multiple dimensional scales
Minkowski Distance	$\left(\sum_{i=1}^{n}\lvert x_i - y_i\rvert^q\right)^{\frac{1}{q}}$	Normed vector space
Chebyshev Distance	$\max(\lvert x_i - y_j\rvert)$	Vector space
Cosine Distance	$\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$	Angle between two points
Bray-Curtis Distance	$1 - 2\frac{\sum_{i=1}^{n} min(x_i, y_i)}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$	Distance between two arrays
Canberra Distance	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{(\lvert x_i + y_i\rvert)}$	Distance between pairs of points

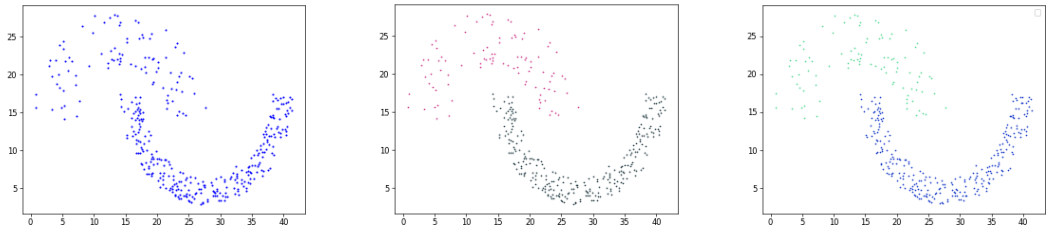
The Euclidean distance is the most common. It computes the square root of the sum of squares of the differences between two objects with n dimensions. In this thesis, the Euclidean distance function is used to calculate the distance between objects and clusters. The Manhattan distance measures the distance along the grid-like layout of city blocks. In the real world, it is impossible to travel indefinitely on a straight line; instead a road network like a chessboard must be used, which is especially common in American cities. The Mahalanobis distance reduces the effect of data distribution. Their variances differ for data with multiple dimensions. In this method, S^{-1} is the inverse covariance matrix, which means that the correlations of dimensions are considered. The Minkowski distance is a generalization method in which q is a positive integer. Different variances can be used for various kinds of distance functions. The Chebyshev distance is a metric in the vector space that is the greatest value in any coordinate dimension. The cosine distance measures the cosine angle between two vectors and ranges from -1 to 1. The vectors could be location information or attributes. The larger the value, the shorter the distance between them. The Bray–Curtis distance calculates the difference between two samples in botany and ecology. The smaller the value, the smaller the difference. The Canberra distance measures the distance between pairs of vectors. When the value is closer to zero, the distance is shorter. The Euclidean distance is widely used and easily understood.

3.2.7Experimental results and analysis

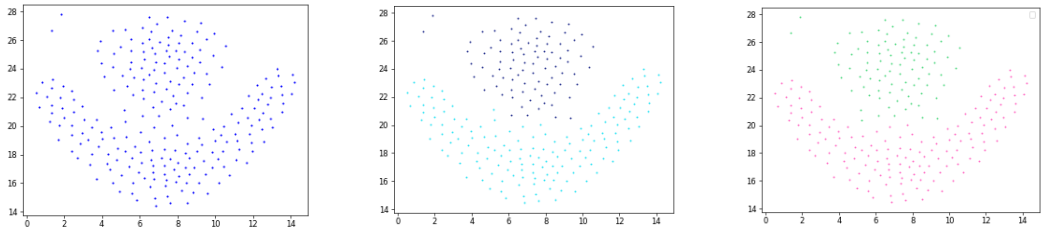
In this section, the six datasets (Fränti & Sieranoja, 2018) shown in Figure 3. 8(a) to Figure 3. 8(f) serve as benchmarks, and the results are compared to verify the proposed method. The first column shows the original data distribution, the second column shows the ground truth of the cluster results, and the third column shows the cluster results using the proposed improved clustering method.



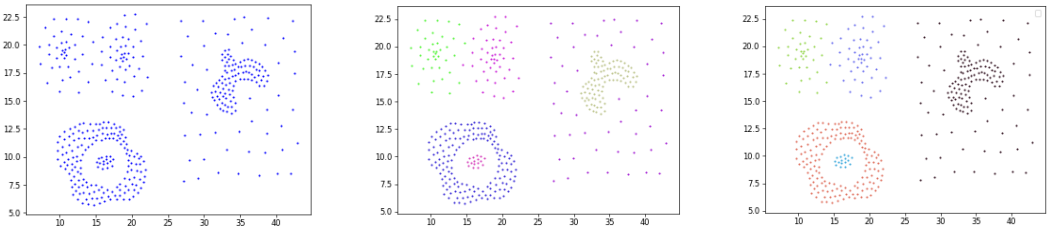
(a) Spiral



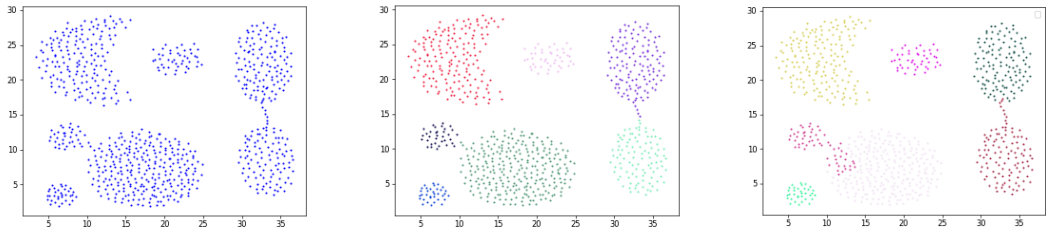
(b) Jain



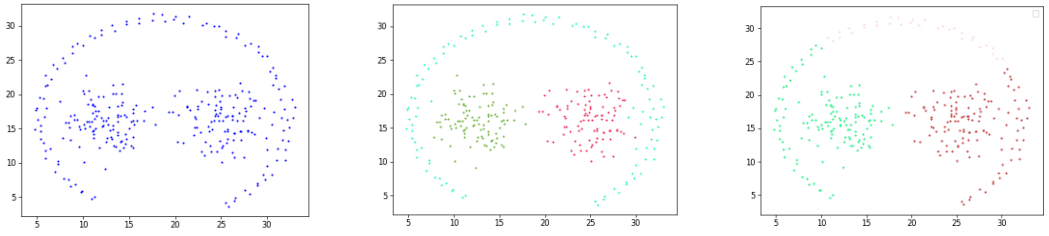
(c) Flame



(d) Compound



(e) Aggregation



(f) Path-based

Figure 3. 8 Dataset and clustering results of benchmark and improved methods. First column is dataset distribution, second column is benchmark of cluster result, and third column is cluster result of the proposed method.

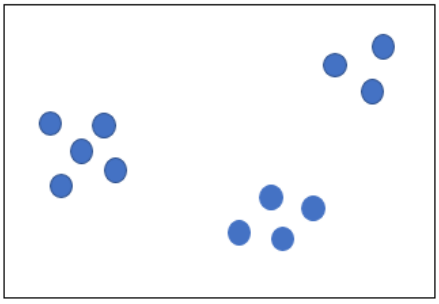
Table 3.3 shows the clustering results with the proposed methods. A comparison of the cluster number and accuracy of the benchmarks and the improved method shows that the latter has great efficiency in detecting clusters. First, the integrated method can correctly detect the cluster number for five of the six datasets. Second, the accuracy reflects the percentage of points that are correctly divided into clusters. Two datasets show accuracy of 100%, meaning that each point is assigned to the correct cluster with the correct number of clusters. Three datasets show accuracy of greater than 85%. Only one dataset shows an accuracy below 80% is path-based dataset. The reason is the distance between two clusters is too small. Similar with compound dataset, there is not clearly space between clusters. When merging the nearest clusters, if the threshold is too small, the number of cluster is large; if the threshold is too large, most of clusters will be merged. It can therefore be concluded that the proposed improved clustering method is highly effective in detecting clusters with multiple shapes and densities.

Table 3. 3 Accuracy of the integrated method

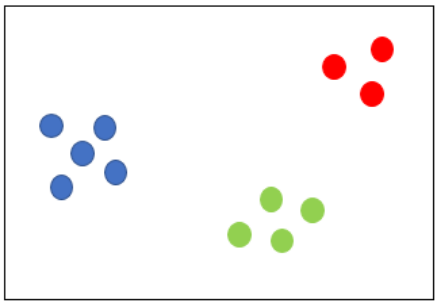
Dataset	Cluster number of benchmark	Cluster number of integrated method	Accuracy of the integrated method
Spiral	3	3	100%
Jain	2	2	100%
Flame	2	2	99.2%
Compound	6	5	87.2%
Aggregation	7	7	95.9%
Path-based	3	3	74.0%

3.3 Voronoi construction

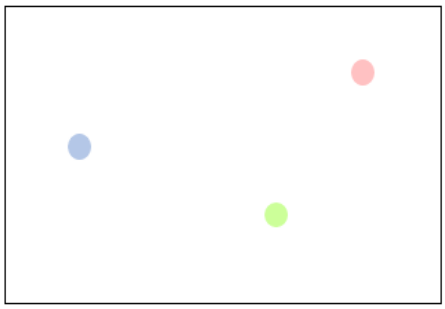
After getting the cluster results of dataset, the center points of clusters are calculated. These center points are regarded as the seed points to construct the Voronoi diagram. Figure 3. 9 shows the process of Voronoi diagram construction.



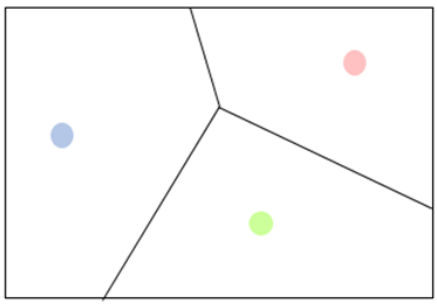
(a) Dataset distribution



(b) Three clusters are identified by using the integrated clustering method



(c) Calculate the center points of clusters



(d) Voronoi diagram construction

Figure 3. 9 The process of Voronoi construction

Three clusters are detected from the dataset. The center points of each cluster are calculated. The perpendicular bisector between two centers points are drawn. The Voronoi diagram is constructed by linking them. This method could be used for region partition by using different kinds of dataset.

3.4 Conclusion

This chapter proposes an efficient Voronoi diagram construction method based on integrated clustering method. The integrated clustering method can detect clusters of multiple shapes and densities. First, key points are selected based on density and distance. Second, after key points are selected, a hierarchical method is adopted to detect clusters. Third, the nearest clusters are merged based on the threshold to form an ideal number of clusters. The method shows great efficiency in detecting clusters with multiple densities and shapes. The method can be used for complex datasets that include clusters of multiple densities and shapes. For example, detection of detailed clusters in GPS data is complicated because the clusters have different densities and shapes. This method can be used to determine clusters with high efficiency. This method has two advantages. First, this method does not require assignment of the number of clusters and cluster center points. The cluster center points have high values for kernel density and are far from each other. Second, only one parameter must be set manually. The center points of each cluster are regarded as seed to construct Voronoi diagram. This method can be used by different kinds of dataset for region partition.

Several improvements must be addressed in future studies. Automatic choice parameter and non-parameter clustering methods remain research problems. Even though the proposed method needs only one parameter, it still based on the users' experience. Noise is another problem that could reduce the effectiveness of this method. We plan to consider improving the ability of noise proofing in complex conditions.

Chapter 4 Analytics of spatial distribution characteristics of the elderly

Chapter 3 described Voronoi construction method based on an integrated clustering method. The integrated clustering method can detect cluster with multiple densities and shapes features. In this chapter, Voronoi construction based on the integrated cluster method will be used for analyzing spatial distribution characteristics of the elderly, together with the further developed new methods. The spatial distribution patterns of the elderly are examined and determined by means of smart card data. First, the spatial distribution of the elderly population in a city is analyzed and presented in Voronoi diagram, which give the detailed description of the elderly living. Voronoi diagram construction method based on the integrated clustering method is used to partition the city into the detailed regions which can better describe the spatial clustering of places where the elderly living. Several assumptions are made to identify the home locations of the elderly, and these are presented in the Voronoi diagram. Second, the spatial distribution pattern of the elderly is measured and explained by a newly proposed PoI-based elderly livability index that is computed based on the determined factors with different weights on the urban facilities needed. The five identified facilities in this study are restaurants, parks, hospitals, shops and bus stops. The numbers and spatial distributions of the five factors for each region are computed using PoI data from Baidu map. Third, the spatial connectivity between the elderly living regions is used to describe where the elderly frequently travels. A quasi-gravity model is developed to reveal the relationship between the connectivity and the PoI-based elderly livability index of the elderly.

4.1 Introduction

In this era of population aging, it is essential to understand the spatial distribution patterns of the elderly in a city and explain the reasons. With the continued improvements in quality of life and subsequent rise in life expectancy, most countries are paying strong attention to reducing the

corresponding negative effects. Several solutions have been proposed, such as gradual or delayed retirement schemes (Burtless, 2013; Kim, 2011), extension of working hours and encouragement for the elderly to form organizations with younger people to learn from each other and share their experiences. Urbanization should create many new job opportunities that attract the elderly working in cities. As indicated above, the study of the elderly distribution patterns has become a research topic of great interest to urban planners and policy makers (Wong et al., 2018).

Two main data sources are commonly used to explore human spatial patterns. The traditional survey questionnaire (Boschmann & Brady, 2013) is the most common method of collecting travel information. The interviewees answer several questions concerning movement behavior, such as travel mode, travel time and personal information. This information could contribute to the understanding of spatial and travel patterns. However, it is impossible to collect large data samples because of labor and time (J. Wang et al., 2019), and location information accuracy cannot be guaranteed. With technological developments, big data have brought new innovations and insights. For example, positioning equipment can easily track location information in real time and provide greater accuracy in mobility pattern analysis. Smart card data record detailed stop locations and times of card holders taking public transport. Not only is a huge volume of spatial information provided, but these data include high-quality location information. However, these data have a number of limitations. First, travel purpose is not very clear. Second, no socioeconomic data are included (Mohamed et al., 2016). In this chapter, the smart card data of the elderly is collected as the dataset to detect spatial patterns. Beijing as one of the most highly developed cities, 24.5% of the population is elderly (Beijing Committee on Aging, 2018). Beijing is thus ideal city for analyzing elderly mobility behavior.

The chapter systemically presents elderly spatial distribution patterns using smart card data. First, the spatial distribution of the elderly population in the city is analyzed and presented in Voronoi diagram, give the precise description of the elderly living. The city is divided into a more detailed Voronoi diagram. Voronoi diagram is a method to partition plane into multiple polygons (Erwig, 2000). The basic idea is to link the perpendicular bisector of two adjacent seeds to have polygons.

In this chapter, the seeds are the centers of clusters. To find these centers, an integrated clustering method is developed based on the total number of elderly both boarding and alighting from bus at each stop. The distributions of the elderly by Voronoi diagram and by administrative regions are compared. Second, the spatial distribution pattern of the elderly is measured and explained by a newly proposed PoI-based elderly livability index. The elderly's residential conditions are considered based on five kinds of PoI data: shops, parks, restaurants, hospitals and bus stops. Finally, the spatial connectivity between the elderly's living regions is used to describe where the elderly travel frequently. A quasi-gravity model is developed to calculate the connectivity of these regions. In this chapter, one-week smart card data of elderly from Beijing Municipal Commission of Transport is used as the data source. The total number of records around 3 million from April 10, 2017 to April 16, 2017.

4.2 Study area and data sources

4.2.1 Spatial distribution of the elderly in Beijing

Beijing is the capital of China, which is located in the north of the country and covers a land area of 16,410.54 km^2. With rapid economic growth, urbanization development and increased resident population, Beijing has become one of the most densely populated cities on the planet. By the end of 2017, the registered resident population of Beijing had reached 13.59 million (Beijing Committee on Aging, 2018). At the same time, the increasing elderly population caused a considerable problem for making suitable urban planning policies for city development. According to statistics data, the resident population of the elderly which is 60 and above is to 3.33 million, which accounted for 24.5% of the total population. Table 4.1 gives the total number of the elderly for each administrative district.

Table 4. 1 Number of the elderly for each administrative district (Unit: ten thousand)

Districts	Population of over 60	Percentage
Dongcheng	27	27.9
Xicheng	38.9	26.9
Chaoyang	57.1	27.2
Fengtai	33.3	29.2
Shijingshan	11	28.9
Haidian	49.5	21
Fangshan	18.2	22.3
Tongzhou	18	23.4
Shunyi	14.3	22.5
Changping	13.4	21.5
Daxing	14.4	20.6
Mentougou	6.6	26.3
Huairou	6.3	22.3
Pinggu	9.2	22.8
Miyun	9.7	22.1
Yanqing	6.4	22.2

Figure 4.1 shows the spatial distribution of the elderly among the 16 administrative districts. The bars show the ratio between the elderly (in gray) and non-elderly (in yellow) in each district. The black circular lines indicate the Second to the Sixth Ring Road. The coloring of the districts indicates the number of the elderly in the district, the darker red district, the more elderly in that district. Chaoyang, Haidian, and Xicheng districts are home to more elderly than any other administrative districts in the city.

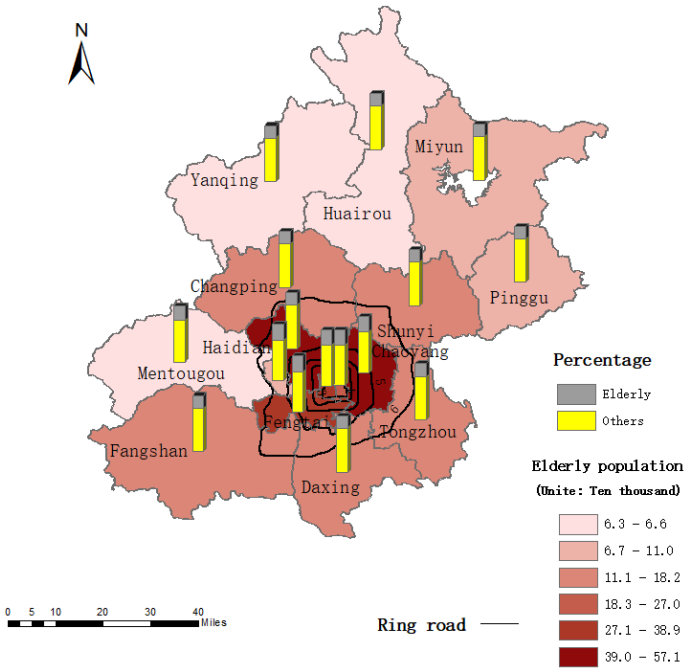


Figure 4. 1 Spatial distribution of the elderly and non-elderly population in Beijing

The number of the elderly increased by 4.2% from 2012 to 2017. The bias regarding the number of the elderly in Beijing is high, and it has been forecast that the population of household registered the elderly will exceed 4 million by 2020. With this trend, the need for adequate care for the city's aging population will become a serious issue for Beijing, and this is likely to be mirrored in the world's other megacities. Hence, it is essential to recognize the importance of and need for a degree of travel independence for the elderly, not simply as an act of kindness, but also because it makes economic sense in that it would contribute to the smooth running of the city and hence, to some degree, better ensure financial equanimity and prosperity. A well-planned city that can cope with an ageing population trend in terms of future mobility needs has undeniable importance.

4.2.2 Smart card data

In this thesis, smart card data comes from Beijing Municipal Commission of Transport is used as the data source. The data is generated when the elderly getting on and taking off bus by swiping their smart cards. Smart card data comprise many information, some of them do not needed in this

thesis. Table 4.2 shows an example of the smart card data. There are three types of smart card data: (a) the general smart card for adults; (b) the smart card for the elderly (i.e., over 60 years); and (c)the smart card for students (primary school, high school, and university students). Because the percentage of student smart card holders is not high and their mobility behavior is relatively stable, it is not discussed in this thesis. The focus of this thesis was on the elderly' mobility behavior as revealed by smart card data analytics. Therefore, the following analysis and discussion considers the smart card data of the following two groups: (a) the adults who also use the smart card, (b) the elderly who used smart the card.

Table 4. 2 Information in smart card data

GRANT_CARD_CODE	ON_STATION_NAME	ON_STATION_TIME	OFF_STATION_NAME	OFF_STATION_TIME	ON_LON	ON_LAT	OFF_LON	OFF_LAT
100075104162 1936	Hepingmendong	2017/4/12 12:38:01	Cuiweilukou	2017/4/12 12:57:20	116.37 9431	39.898 837	116.29 4938	39.906 309
100075104162 2646	Nanwu	2017/4/12 18:49:01	Xipingzhuang	2017/4/12 19:03:30	116.26 8963	39.958 863	116.22 79	39.953 724
100075104162 4088	Zhengchangzhuang	2017/4/12 8:02:01	Wuzhuang	2017/4/12 8:37:08	116.26 8362	39.886 734	116.22 4165	39.891 885
100075104162 7104	Fengcun	2017/4/12 15:19:01	Shichangcun	2017/4/12 15:32:37	116.10 553	39.909 198	116.09 3205	39.886 518
…	…	…	…	…	…	…	…	…

Seven kinds of attributes are selected to be used in this thesis, they are the card ID, departure stop name, departure stop longitude and latitude information, departure time, arrival stop name, arrival stop longitude and latitude information, arrival time. Each smart card has unique card ID for identification, each ID stands for one elderly person. Smart car ID is an important attribute in the following analysis. One elderly person could have many trips in one day or several days. Departure and arrival stop names are similar with smart card ID which are used to identify the stops. As the elderly must swiping cards twice contain get on and take off bus. The data record the time when

swiping card and location information of bus stops. The location information of bus stop is made up by longitude and latitude of stop. Figure 4.2 shows the distribution of public bus stops and illustrates the multiple payment methods.

Figure 4. 2 Bus stop distribution and multiple payment methods (Source of right sub-figures: Baidu images)

To demonstrate why smart card data is used for this thesis, we first compared the smart card data and the questionnaire data, taking the information in the questionnaire for the fourth travel survey of Beijing as an example. Table 4.3 shows the differences between the smart card data and the data from the questionnaire.

Table 4. 3 A comparison between smart card data and questionnaire

	Smart card data	**Questionnaire data**
Updating frequency	Hourly for whole year round	Every 5 years
Sample size	5 million daily	10 thousand each time
Time period that data reflect the situation in the real world	For whole year round with one hour time interval	For one day in the five years
Personal Information	Cardholder types: • Student • Elderly • Adults	Interviewee: • Age • Gender • Job • Home location • Family member • Income
Travel purpose	Not available directly from the data, but can be speculated roughly based on the developed algorithms	Clearly described
Location of departure and arrival bus stop	Precise bus stop location with longitude and latitude coordinates	General description of the location without detained coordinate information
Departure and arrival time	Precise time information with second level of accuracy	Estimated approximation with hour level of accuracy
Travel mode	Clearly recorded for each trip	Described for different trips
Trip with interchange	Clearly recorded for each trip	Described for each trip
Travel Frequency	Travel frequency is precisely recorded daily for the whole year round	Travel frequency is described for the date of survey

Compared with the travel survey, the smart card data have several unique advantages and inevitable disadvantages. On the one hand, the smart card data can provide near–real time information in 1 hour, a large sample size up to 5 million (versus 10,000 for the questionnaire), precise location

information with latitude and longitude coordinates, and more frequent information at an hourly rate (compared with the 5-year interval of the questionnaire). On the other hand, the data from the questionnaire provide clear information on the interviewee background and details of the travel purpose, which are not directly available from the smart card data, although it can be speculated to a certain degree with the use of developed algorithms for smart card data analytics.

The advantage of the smart card data for Beijing is that the elderly should swipe their cards twice, when getting on and off the bus. The detailed information for the departure and arrival stop location and time can be obtained. However, some cities, such as Hong Kong, do not require the card to be swiped when getting off the bus.

4.2.3 Public transportation systems

The bus and subway systems are Beijing's two main public transportation modes, and 20 subway lines are currently in operation. Most of the elderly in the city take the bus for their daily trips rather than using the subway. Most subway lines are run by the Beijing Subway Company, with a few runs by the Beijing MTR Company (the latter being related to the Hong Kong MTR). The Beijing Bus Public Transport Company, with 860 scheduled bus services and 406 nonscheduled bus services, is responsible for ground transportation services.

4.3 Data preprocessing and distribution functions

4.3.1 Data cleaning

Smart card data records trip information from two transportation modes: the bus and the subway. The first data cleaning stage is to remove any unnecessary information recorded in the raw smart card data that would not be used in later analysis. For instance, the Vehicle_ID and City_Name will be unnecessary for the later mobility behavior analysis, and the corresponding analysis at a later

stage will be more efficient without the redundant data. It is possible that some data are recorded twice because of system error, and these data must also be removed.

The second data cleaning stage is to detect and remove errors in the raw smart card data. For example, some records report a travel frequency of more than 30 times per day. Given the nature of the elderly, this is obviously an error. This type of outlier was removed by simply setting the maximum travel frequency to 6 as the criteria, for example, and by removing records with a travel frequency of 7 or more per day. Another example is provided by a record detailing that the travel duration exceeded 3 hours. This is also regarded as an outlier, and the information for such a trip would be deleted from the records as an outlier or gross error. The rationale for this is that the running time of a whole bus line is normally no longer than 2 hours. To be safe, trip records with a travel duration longer than 1 hours are justifiably removed; although some bus lines could be long and one line may encounter a traffic jam, people seldom ride from the first stop to the last stop. In an extreme case, one may travel more than 10 times a day by bus, or the travel may have a duration of longer than 3 hours, but the number of such cases is low, and likewise, these data have little effect on the analysis results. In this thesis, we removed items that fulfilled the following conditions: (a) travel distance longer than 16 km, (b) travel duration longer than 60 minutes, and (c) travel frequency greater than 6 per day.

4.3.2 Data quality analysis

The smart card data quality was improved after data cleaning and the removal of outliers in the data sets. Table 4.4 shows a comparison of the trip records before and after the data cleaning process. For example, 99.7% and 38.8% of the original trip records are retained for the elderly and adults, before removing trip records that may contain errors. Statistically, this sample size is sufficient for the estimation of the population. In fact, this is the power of big data: there will still be enough samples for big data analytics even after the suspicious records are removed.

In the final row of Table 4.4, “Travel by Bus after Data Cleaning” the indication is that a card holder

takes the bus to travel. It is seen after data cleaning that the record for the elderly is 82.5% of the overall travel by bus. Therefore, it is reasonable to use the samples of “Travel by Bus after Data Cleaning” to study the elderly mobility.

Unsurprisingly, most the elderly prefers to avoid taking a bus with interchange and taking the subway for safety reasons. Others prefer to use the subway because it is usually faster, but it requires them to use stairs. Hence, it is clear that the elderly prefers to take a bus without transferring to the subway, due to safety concerns. For instance, it may be dangerous for the elderly to climb stairs after taking a subway. The elderly’s bus travel is discounted, whereas the use of subways is not discounted. Only 1835 trips were made by subway, accounting for only 0.3% of the total number, and 99.7% of trips were made by bus. For this reason, we only consider the trips made by bus in this thesis.

It is clearly that the main public transport mode for the elderly in Beijing is bus. It is assumed that the elderly takes one individual trip each time, thus avoiding the need to change bus lines or interchange to other public transport modes, such as the subway.

Table 4. 4 Smart card data before and after data cleaning

	Elderly		Adults	
	Number	%	Number	%
Raw Data of Travel by Bus	533,477	99.7	3,298,413	38.8
Raw Data of Travel by Subway	1,835	0.3	5,198,003	61.8
Travel by Bus After Data Cleaning	441,560	82.5	2,869,493	33.8

4.4 Methodologies

This chapter developed a series of data-driven methods on spatial distribution of the elderly aiming to answer the following three questions: a) Where do they live? b) Why the elderly distribution like this? c) Where do they go frequently? The series of methods comprise the framework for the elderly distribution pattern, as shown in Figure 4.3. The question "Where do they live?" is answered by "Elderly Population Distribution based on the Voronoi Diagram" and "Identify Home Location of Each Card Holder". The question "Why do they live in there?" is answered by "The PoI-based Elderly Livability Index of Each Polygon in the Voronoi Diagram". The question "Where do they go frequently?" is answered by "Spatial Connectivity between the Elderly Distribution Regions". Two data sources are used in the data-driven methods: a) smart card data and b) PoI data. Both were described in the previous section.

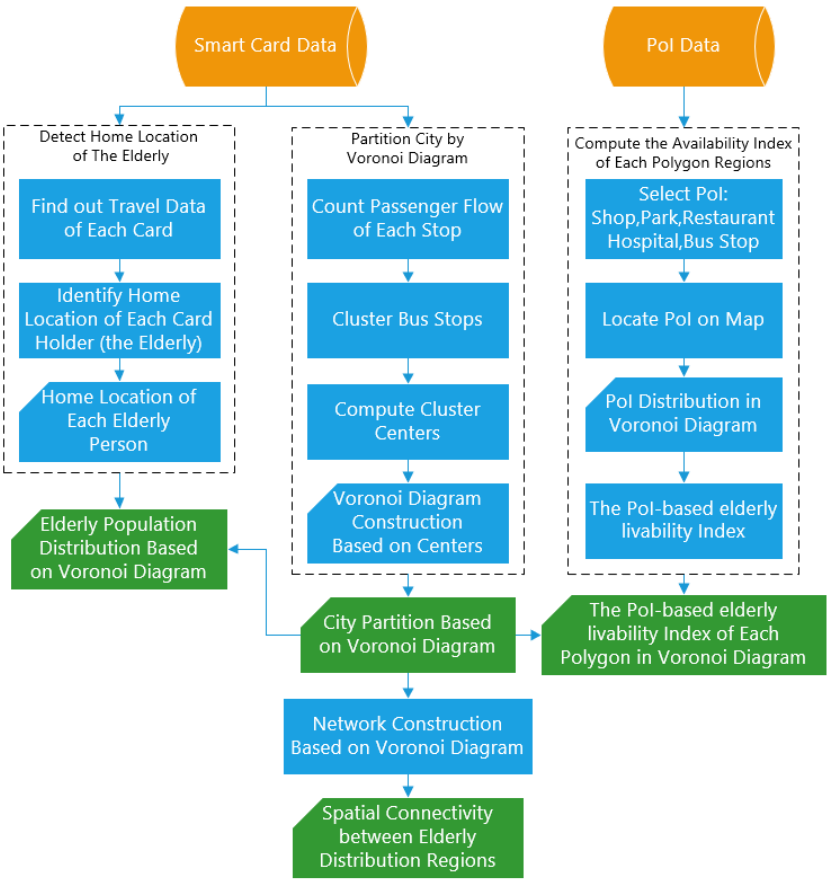


Figure 4. 3 Analytics framework of the spatial distribution patterns of the elderly

4.4.1 Spatial distribution of the elderly population

The methods for identifying the spatial distribution of the elderly will be developed to answer the question: “Where do they live”. Here we need to develop approaches for the following two tasks: what is more suitable way to partition the residence clusters of the elderly in a city and how to determine the home location of the elderly.

A method of city partitioning for the elderly residential regions based on Voronoi diagram

To describe the spatial distribution of the elderly, we require a method to properly partition a city into the residential regions of the elderly. Traditionally, administrative regions, such as the districts of a city, are used as the basic geographic unit for censuses. However, in many cases, the spatial distribution of the residential areas may not exactly follow the outline of the administrative regions. For example, a residential region may cross two administrative regions, leading to difficulties regarding the follow-up statistics. Therefore, a more natural and proper method of city partitioning for elderly residential regions must be developed. Here, we propose a method of city partitioning for the elderly residential regions based on the Voronoi diagram as an alternative solution. This method adopted the integrated clustering method which is proposed in the chapter 3.

The Voronoi diagram partitions a plane into regions close to given sets of seed points (Fortune, 1987). In the thesis, the plane refers to the whole area of a city, the seed points are the centers of the elderly’s residential areas, and the Voronoi diagram is the partition of a city where each polygon represents a cluster of the elderly residences (Shao et al., 2019). The logic flow of the proposed method, which include four steps of city partitioning for the elderly’s residential regions based on Voronoi diagram is illustrated by the middle column of Figure 4.3.

- First, counting the elderly passenger flow which are total number of getting on and taking off bus of each stop. Sorting all stops $S_N = \{S_1, S_2, \dots S_n\}$ in descending order based on the number.
- Second, a density-based clustering method is developed for clustering bus stops. The principle

of this clustering method is depicted by the pseudo code in Table 4.5. A distance threshold value r is set in advanced. According to the sorted result in the first step, for each stop S_i, the stop S_j is identified, where stop S_j has the shortest distance with stop S_i among all the bus stops. If the distance is smaller than the threshold, then S_i and S_j belongs to the same cluster, while if it is larger than the threshold, the stop S_i is regarded as a new cluster center. Continue such process until all the bus stop are clustered.

- Third, after all the bus stops are cluster into different clusters the center point of each cluster is computed, and the arithmetic mean is used for computing the coordinates of each center.
- Finally, the center points of each cluster are regarded as the seed point, linking the perpendicular bisector of two adjacent seed points, the polygons of Voronoi diagram is thus generated. We choice Arcgis 10.2 (Johnston et al., 2001) to generate the Voronoi diagram by importing the result of center points.

Table 4. 5 Pseudo code of developed clustering method

Algorithm: Clustering method based on high frequency and distance
Data: Smart card data S
Input: Distance threshold r
Cluster indicator $c = 0$
Output: Cluster number
Set of cluster C_n
Count passage flow of each stop and sort these by descent based on the number $S = \{S_0, S_1, S_2 \dots S_n\}$
For i to S.length
If i = 0 then
Stop S_i is new cluster core, insert c into C_n
$c++$
Else
Let $d[\,]$ is new array
For j in $i-1$ do
$d[d_{ji}] \leftarrow distance(S_j, S_i)$
Sort $d[(d_{ji})_0, (d_{ji})_1, \dots\,]$ by ascending
If $(d_{ji})_0 < r$
Stop S_i and S_j are in the same cluster, insert C_j into C_n
Else
Stop S_i is new cluster core, insert c into C_n
$c++$

Voronoi diagram-based method for partition city into residential regions developed in this research is more rational and reliable. This method considers not only the spatial setting of the bus stops in a city but also the number of residents who use the bus stops. Compared with the partitioning of a city by administrative regions, the proposed method has the following two advantages: a) it is based on the natural clustering of residents and avoids the circumstance that a residential region may cross two administrative regions, and b) it can solve the false address registration problem whereby one person officially registers in one administrative region but frequently lives in another administrative region. The data-driven method based on the usage of bus stops reflects the real/frequent living place. The proposed method in this chapter, which is based on these data, is therefore more reliable.

A method of identifying home location of the elderly

In this thesis, a method is proposed of using smart card data to identify the home location of the elderly based on the locations of the bus stops they use frequently. We make the following four assumptions.

Assumption (a) Home location is close to the most-used bus stop

It is possible to estimate the home locations of the elderly using the locations of the most frequently used bus stop. Normally, for a commuter, home and work are the two most frequent places in the daily itinerary of trips. Even though the elderly does not have a regular workplace, the home location is a high travel frequency origin and destination.

It is highly likely that the home location of an elderly person is close to the most frequently used bus stop. More than 50% of person will return to stop which is located within 1 km of their departure stop of day (Chakirov & Erath, 2012). Normally, the walking distance of the elderly is about 400m (Zhou, Shen, & Jiang, 2016). In general, the stop should be within walking distance from home.

Some exceptions could exist in practice, although the percentage of such cases is not likely to be very high. For example, one may use different transportation modes, such as the subway, to finish the rest of trip. This is most likely when the bus stop is not near the home location of the traveler.

However, the rate of interchange and changing from another public transport mode to the bus is very low. For example, in Beijing, only 0.3% interchanged and changed their transport mode according to 502165 records of smart card data from April 12, 2017.

Assumption (b) The bus stop of earliest departure or latest arrival during an elderly person's daily trip has a very high chance of being the home location. Barry et al. (2002) proposed similar assumptions for estimating origin-destination trips, such as home-office trip.

Assumption (c) The bus stop of earliest departure for all the recorded days has the highest probability of being the home location. Based on these assumptions b) and c), frequency and time of day are the two main factors used for identifying bus stops regarding as the home location.

Assumption (d) One elderly person uses one smart card mainly for his/her trip.

Figure 4.4 shows the logical flow of the method for identifying the home location of the elderly. The logic follow of the proposed method is as follows:

- First, we filtered the dataset, listed all smart card IDs and confirmed that each corresponded to a different elderly person.
- Second, we counted the total number of times that a stop was the first departure or last arrival stop each day. A minimum threshold α was set to four to filter the stops because most elderly are unable to take public transportation with a high frequency.
- Third, if the total number of instances of any such stop was higher than the threshold, the home location of the smart card ID could be detected; otherwise, it could not be detected.
- Fourth, if only one stop had a total number of instances larger than the threshold, the stop was identified as the home location of the smart card ID. If two or more stops met this criterion, the time of day was considered as the deciding factor in the second step. We sorted the stops by the time of day, and the stop used at the earliest time was considered the home location.

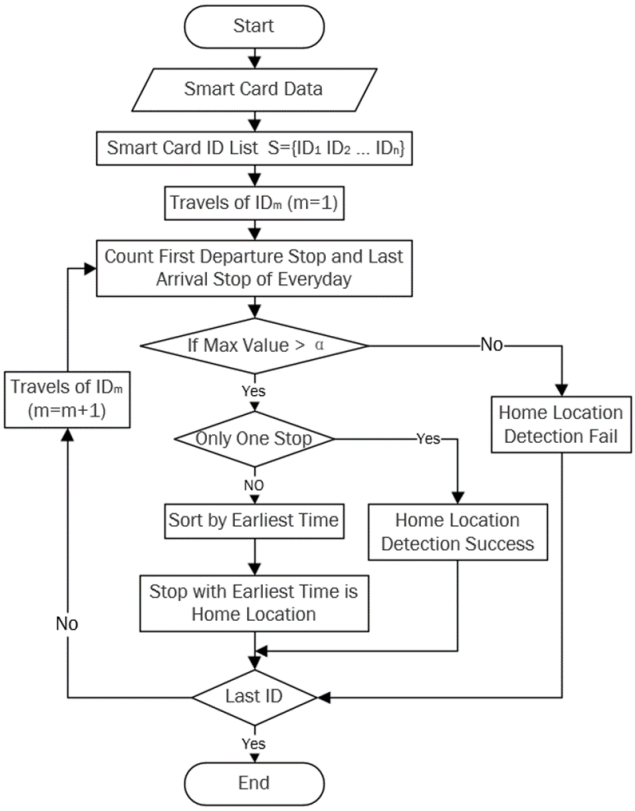


Figure 4. 4 Logical flow of the proposed method for identifying home location of the elderly

The advantages of this home location identification method are that the home location detected reflects the latest and regular home location of an elderly person if his/her registered home location is not updated or if he/she has more than one. The accuracy of the detected home location of the elderly would be much higher if accumulated smart card data is available for a longer period, such as several months.

Based on the above assumptions of the proposed method for home location identification, the home locations of most of the elderly card holders were identified using the smart card data. However, a minority could not be identified. In most of these cases, the total number of times that the elderly traveled from any bus stop was very low, or even zero, according to the smart card data. This is one limitation of the proposed method for elderly home location identification due to the data quality of the smart card data. Fortunately, this small number of missing locations had no substantial effect on the overall spatial distribution analysis of the elderly population of the city and can thus be ignored.

Based on our assumptions, the bus stop should be within walking distance of home. Some exceptions could exist in practice, although the percentage of such cases is not likely to be very high. For example, one may use other transportation modes from home, such as the subway, and then change to a bus for a subsequent trip. This is most likely when the bus stop is not near the home location of the traveler. However, in general, it is reasonable to assume that the departure bus stop is close to the home location of an elderly person because the rate of changing from another public transportation mode to the bus is very low, as indicated by the Beijing data.

Moran's I for determining the city partitioning methods for representing the elderly spatial distribution

In this chapter the appropriate city partitioning methods for representing the spatial distribution of the elderly in a city was determined using the spatial correlation defined by Moran's index (Moran's l). According to the first law of geography, which was proposed by Waldo Tobler, is "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Spatial autocorrelation is an important concept that is widely used to measure the correlation among variables in geographical space, and in this study Moran's I was used for analyzing the correlation among regions in terms of the elderly distribution on different scales. Moran's I is defined as follows:

$$I = \frac{n \sum_i \sum_j w_{ij} (x_i - \bar{x}) \left(x_j - \bar{x}\right)}{\sum_i \sum_j w_{ij} \sum_i (x_i - \bar{x})^2} \tag{4.1}$$

where n is the number of geographical units, x_i and x_j are the attributes of units i and j, w_{ij} is the weight between them. Values for Moran's I range from -1 to 1, where positive values mean the feature distribution is clustered, negative values mean the feature distribution is dispersed, and zero means there is no spatial autocorrelation.

In this chapter, the spatial correlation derived from Moran's I was used to determine the appropriate city partitioning method for the elderly population distribution. In the following case study of Beijing, first, the city was spatially divided into regions on two methods: a) 16 administrative regions and b) a 147-polygon Voronoi diagram. Second, the spatial distribution of the elderly

according to the three methods was derived. Third, Moran's I value for each of the method representations of the elderly distribution was computed. Finally, we evaluated which scale was the most appropriate for representing the spatial distribution of the elderly. By comparison of the Moran's I values for the three scales of representation, the Voronoi diagram was found to the most appropriate division of the elderly population distribution.

4.4.2 The model of PoI-based elderly livability index

To answer the question "why the elderly distribution like this?" this research proposes the model of the PoI-based elderly livability index. PoI-based elderly livability is an indicator of whether a region is suitable for the elderly to live. The higher the index value, the more suitable the region is. In this chapter, the model of the PoI-based elderly livability index is designed to be computable based on PoI data, which is open and widely available from the Internet.

The model includes two parts: a) selecting the factors most strongly related to the PoI-based elderly livability index in terms of PoI and b) determining the quantitative relationship between the PoI-based elderly livability index and the related factors.

Determine the factors related to the PoI-based elderly livability index by Pearson correlation coefficient

In this chapter, the factors related to the PoI-based elderly livability index are determined using the Pearson correlation coefficient, which is widely used to measure the relationship between two dependent variables, and can be defined as in Equation (4.2):

$$r_{xy} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2}\sqrt{(y_i - \bar{y})^2}} \tag{4.2}$$

where $\bar{x}$ and $\bar{y}$ are the means of x and y. The coefficient r_{xy} ranges from -1 to 1, where a positive value indicates that the variables are directly related, a negative value indicates that the

variables are inversely related, and a value of zero means that they are uncorrelated.

If a region has a large elderly population, it can potentially be regarded as suitable for the elderly to live. This is based on two considerations. First, if people are free to choose places to live in a city, people tend to choose the most suitable places, and this is also applicable to the elderly. Second, if a region has a large population, facilities related to the elderly will be enhanced either by the government through urban planning (such as hospitals) or driven by the commercial market (such as shops).

In this chapter, we hypothesized that the elderly distribution would have a strong relationship with the distribution of a number of PoIs. To evaluate this assumption and identify the factors related to livability, the Pearson correlation coefficients between the elderly population and PoI data of each polygon in the Voronoi diagram are calculated. PoI with a positive correlation and a high value for the elderly population distribution in the Voronoi diagram are chosen as the element factors for the PoI-based elderly liability index model.

The PoI-based elderly livability index model

Through the above correlation coefficients of the elderly population distribution and PoI, the following five kinds of PoI were identified for computing the PoI-based elderly livability index: shops, parks, restaurants, hospitals and bus stops. The index is quantified by this set of factors with an emphasis on public facilities, which are extremely important to the elderly in daily life, and these facilities are indicated by the spatial distribution of corresponding PoI. The model of the PoI-based elderly liability index is thus formulated as follows:

$$L\_Index_i = w_s * p_{si} + w_p * p_{pi} + w_f * p_{fi} + w_h * p_{hi} + w_t * p_{ti} \quad (i = 1,2,3,\dots) \qquad (4.3)$$

where L_Index_i is the PoI-based elderly livability index; w_s, w_p, w_f, w_h and w_t are the weights of the shop, park, restaurant, hospital and bus stop, respectively; p_{si}, p_{pi}, p_{fi}, p_{hi} and p_{ti} are the respective numbers of the five kinds of PoI in each region; and i is the ID of each

region.

The weights of the five kinds of PoI are determined based on three considerations: a) the correlation coefficient analysis results, b) elderly citizens' needs and c) the travel purposes of the elderly, such as travel purpose of the elderly in Beijing by Xia and Guan (2013). In this chapter, we set the weights of the shop, park, restaurant, hospital and bus stop parameters as 0.21, 0.41, 0.15, 0.11 and 0.12, respectively, based on the above three considerations.

The PoI-based elderly livability index proposed in this chapter can explain why the elderly distribution like this, and where is a suitable place for the elderly to live. This index can be used to explain the spatial distribution of the elderly in a city. For an individual need point of view, the index can support an elderly person in selecting a region in which to rent or buy a house. The index model can also support the government's urban planning of the construction of suitable facilities for the elderly and corresponding land use.

The model of the PoI-based elderly livability index provides a quantitative method of analyzing how suitable a place is to live. The idea is generic and the index is computationally achievable because it is a data-driven solution based on PoI, which is widely available on the Internet. However, the PoI-based elderly livability situation may be different from one city or country to another. The factors and corresponding weight of the PoI-based elderly livability index model are modifiable. For the example of Beijing, a high proportion of trips taken by the elderly are for physical exercise, and therefore, the weight of the park parameter is larger than the others. For example, in Europe and the U.S., the main travel purpose of the elderly may be for shopping.

The proposed model of the PoI-based elderly livability index in this chapter emphasizes facilities, including shops, parks, restaurants, hospitals and bus stops. However, the PoI-based elderly livability of the elderly may also include social, economic and cultural factors. For example, one might consider a neighborhood's social and cultural background when choosing a place to live. The proposed PoI-based elderly livability index model can be further extended in this direction in the

future.

4.4.3 Connectivity of the elderly living regions

To answer the question "where do the elderly go from one region to another frequently" is researched and answered using connectivity analysis in this chapter. Strong connectivity between two regions where the elderly live indicates more frequent travel between the two. A network is constructed for the analysis. The nodes are the centers of the regions where the elderly live, as represented by the polygons in the Voronoi diagram, and the edges are the connections between the nodes.

Connectivity between the regions of the elderly living

Connectivity is a spatial distribution pattern that describes the connection and its strength between two nodes in a network. In this chapter, network analysis was used to study the connection between regions, such as between the home and shopping. A network was constructed based on the Voronoi diagram. Each polygon is regarded as a node N_i, and the centroid coordinates of each polygon (c_i, c_j) are regarded as the spatial location of the nodes. A directed edge E_{ij} between two nodes (N_i, N_j) is constructed if there is a pair of trips between these areas. Each node may contain a loop L_i when several trips happen in the same polygon. Each edge has weight W_{ij}, which is measured by the total number of trips between node N_i and N_j. Each loop has weight W_i, which is computed as the total number of trips within the unit. We represent a weighed graph $G = (N, E, L, W)$ based on the Voronoi diagram. We calculated three network indicators, namely degree, strength and density, to further understand the elderly spatial connectivity. The degree of each node is the number of other nodes to which it is connected, defined as Equation (4.4):

$$d_i = \sum\nolimits_{j \in N} a_{ij} \tag{4.4}$$

where d_i is the degree of node i; a_{ij} is equal to 1 if node i and j are connected, and otherwise

is equal to 0; and N is the node set of the network. We calculated the degree as well as in-degree and out-degree of each polygon. Finally, the average degree of the network was calculated. The definitions are as Equations (4.5) to (4.7):

$$d_{i1} = \sum_{j \in N} b_{ij} \tag{4.5}$$

$$d_{i2} = \sum_{j \in N} c_{ij} \tag{4.6}$$

$$d_{ave} = \frac{\sum d_i}{n} \tag{4.7}$$

The in-degree is the number of head ends adjacent to a given node. Here, d_i is the in-degree of node i; b_{ij} is equal to 1 if node j head-ends node i, and otherwise is equal to 0. The out-degree is the number of tail-ends adjacent to a given node. Here, d_i is the out-degree of node i; c_{ij} is equal to 1 if node j tail-ends node i, and otherwise is equal to 0. The average degree is the average value of all of the nodes in the graph, and is used to describe the connectivity of the graph as a whole. Here, d_i is the degree of node i, and n is the total number of nodes. In this chapter, the degree was computed to determine if there is connectivity between pairs of polygons in the Voronoi diagram. We calculated the degree of each pair of regions between which travel occurred and sorted these degree values from high to low. High values mean the regions are connected with a large number of other regions, and vice versa.

In network theory, the connectivity strength is the strength of each node equals the count of other nodes connected to it (Zhao et al., 2018). Each pair of nodes can include many connections between them. For example, many elderly citizens can travel between any two regions. Similar to degree, strength is divided into in-strength and out-strength. Average strength is calculated as the average value of strength of all nodes in the graph. The equations (Equations (4.8) to (4.11)) are shown as follows:

$$s_i = \sum_{j \in N} sa_{ij} \tag{4.8}$$

$$s_{i1} = \sum_{j \in N} si_{ij} \tag{4.9}$$

$$s_{i2} = \sum_{j \in N} so_{ij} \tag{4.10}$$

$$s_{ave} = \frac{\sum s_i}{n} \tag{4.11}$$

where s_i is the strength of region i, sa_{ij} is the connection between region i and j, s_{i1} and s_{i2} are the in-strength and out-strength of region i and s_{ave} is the average strength of the network. Strength is an important indicator in this chapter. We regarded the total number of trips between two regions as the measure of strength. The greater the strength, the stronger the connection between two regions.

The density of a graph measures the sparseness and denseness according to the number of edges, and is defined as Equation (4.12):

$$D = \frac{|E|}{|N|(|N|-1)} \tag{4.12}$$

where E is the number of edges and N is the number of nodes. The maximum value of graph density is 1 and the minimum is 0. In this chapter, density indicates the balance of the connectivity of regions, in terms of travels by the elderly, over the whole network. If the value is small, only some pairs of regions have strong connectivity. If the value is large, most pairs of regions have strong connectivity.

In this chapter, density is used to indicate the balance of the connectivity of regions, in terms of travel by the elderly, over the whole network. If the value is small, then only some pairs of regions have strong connectivity. If the value is large, then most pairs of regions have strong connectivity.

A quasi-gravity model based on the PoI-based elderly livability index

A gravity model is commonly used to quantify the relationship between different objects. The gravity model can be expressed as Equation (4.13):

$$I_{ij} = G\frac{m_i m_j}{d^{\gamma}} \tag{4.13}$$

where G is a constant parameter that can be determined according to the problem concerned, m_i and m_j are the mass of object i and object j respectively, d is the distance between i and j., and γ is the order of distance. Many types of distance can be chosen, such as Euclidean distance or Chebyshev distance, and an appropriate type should be determined for the particular application concerned.

In this chapter, we discovered that the relationship between the PoI-based elderly livability index and the connectivity follows a quasi-gravity model, represented in formula (4.14). Here the PoI-based elderly livability index is quantified by that of an elderly living region represented by the polygons in the Voronoi diagram on the elderly population distribution, and the connectivity strength is quantified by that of the edge connecting the two regions.

$$S_{ij} = G\frac{L\_Index_i\, L\_Index_j}{D^{\gamma}} \quad (i \neq j, i = 0,1,2,\dots, j = 0,1,2,\dots) \tag{4.14}$$

where S_{ij} is the strength between region i and j, G is a constant parameter, L_Index_i and L_Index_j are the PoI-based elderly livability indices of regions i and j, D is the Euclidean distance between the centroids of regions i and j, and γ is the order of distance. This quasi-gravity model was verified based on the smart card data sets of the study area.

By using the quasi-gravity model the connectivity strength, we can derive connection strength between two regions based on the PoI-based elderly livability index value of the two regions. Based on this result, we can estimate where the elderly group travel to more frequently from the region

they live.

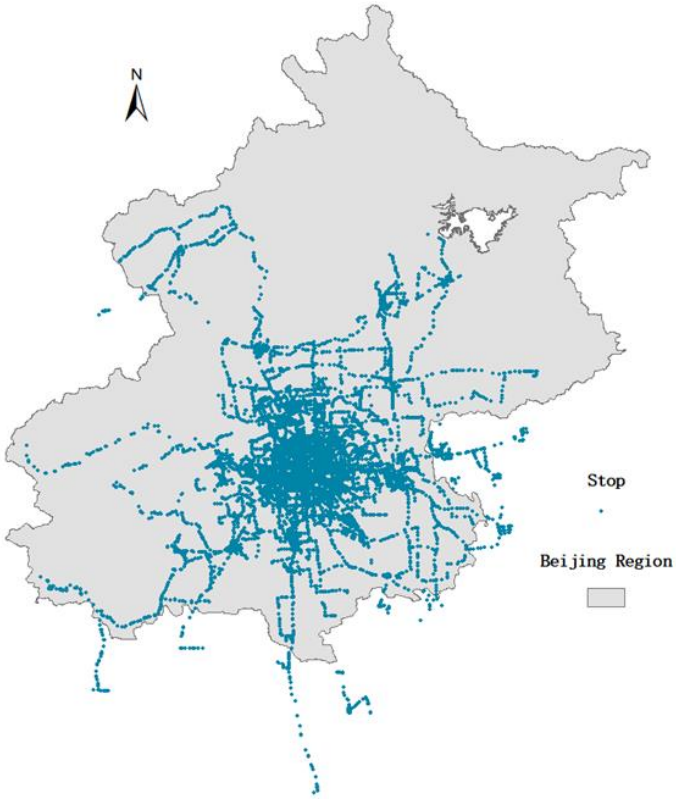
4.5 Application study and results analyses

Beijing was identified using an application study in this research. The developed methods described above were applied to the study using the collected smart card data, bus stop data, PoI data and other related data for Beijing.

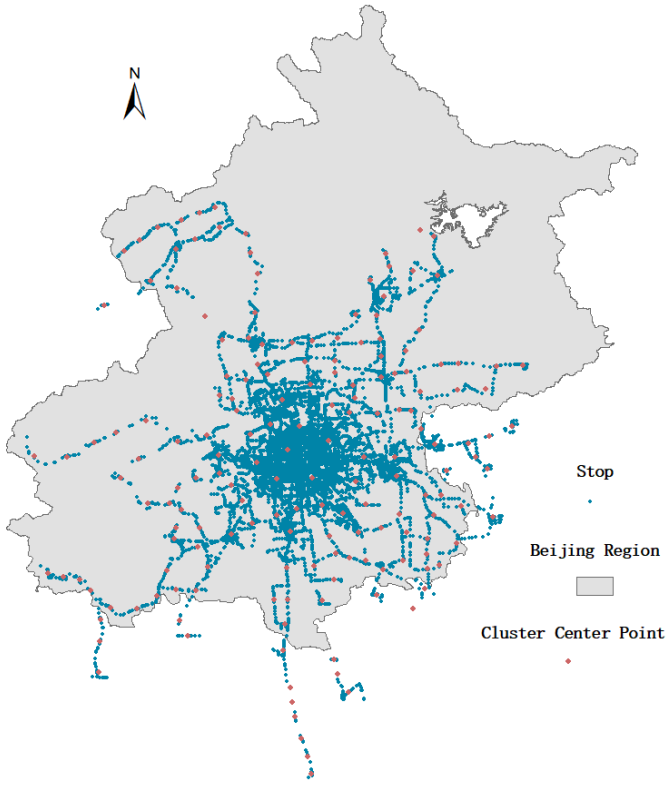
4.5.1 Spatial distribution pattern of the elderly

Spatial distribution regions of the elderly constructed by Voronoi diagram

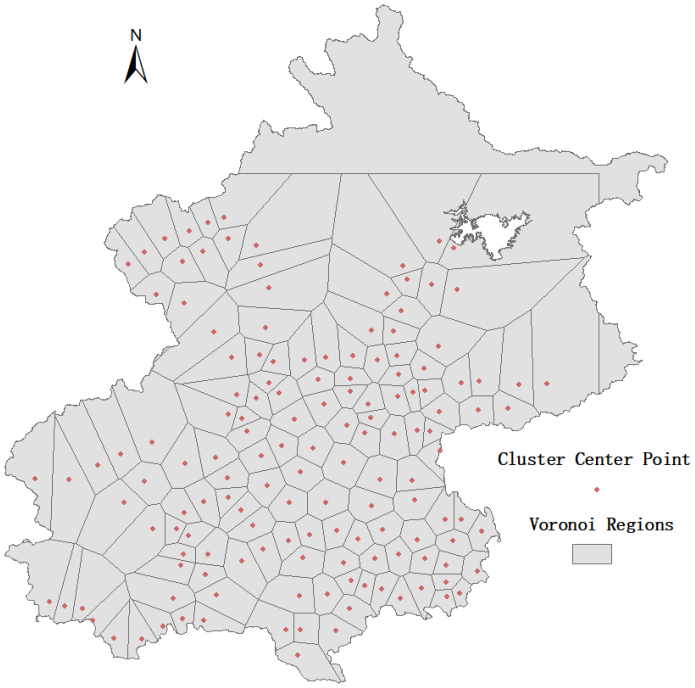
Voronoi diagram was used to partition the city into spatial distribution regions of the elderly based on smart card data. The developed clustering method described above was first used to find the centers of the clusters, which served as seed points. Voronoi diagram was constructed based on these seed points, and the spatial distribution regions were then constructed as a new way to partition the city. Figure 4.5 (a) shows the bus stop distribution. Several bus lines running between Beijing and Hebei province. Some stops located out of Beijing regions. In Figure 4.5 (b), the red points are the cluster center points which are detected by using the clustering method. Figure 4.5 (c) shows the cluster center points and Voronoi diagram. The cluster center point which located out of Beijing regions are deleted. Figure 4.5 (d) presents the Voronoi diagram with the regions as grey polygons, and the blue lines show the Ring Road of Beijing. Because of the limitations of the dataset, in remote areas with a low density of bus stops or none at all, the corresponding polygons are much larger than those of areas with a high density of bus stops and elderly residents. In Figure 4.5, for example, there are large polygons in the northern part of the city, and most of the small-polygon regions are located in the center of the city. This indicates that most of the elderly live in the center, and few in remote areas. The reason may be that many facilities needed by the elderly for daily life are located in these central regions. This result is also indicated by the spatial distribution of the PoI-based elderly livability index later in this chapter.


N
Stop
Beijing Region

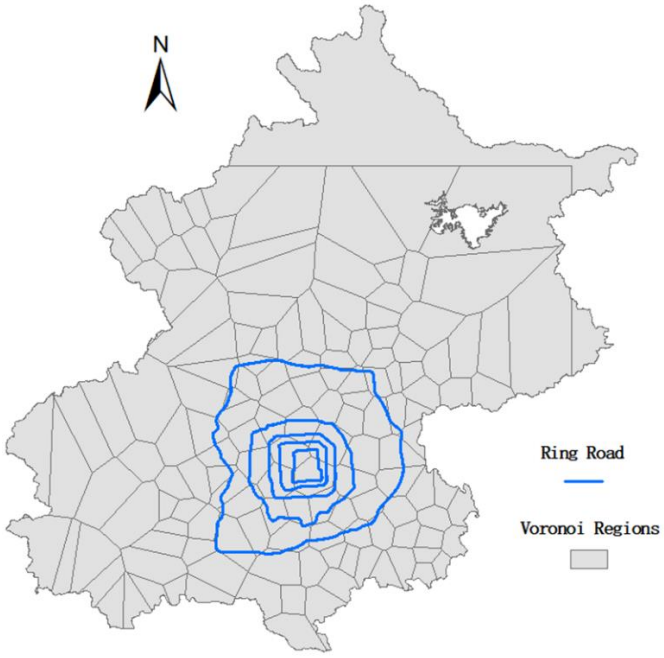
(a)


N
Stop
Beijing Region
Cluster Center Point

(b)



(c)



(d)

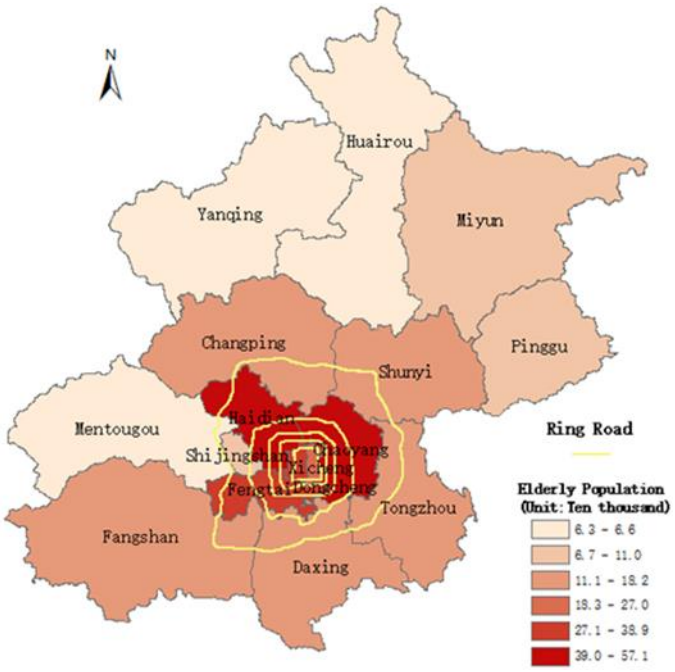
Figure 4. 5 (a) Spatial distribution of bus stop; (b) Spatial distribution of bus stop and cluster center; (c) Spatial distribution of cluster center and corresponding polygons in Voronoi diagram in Beijing;

and (d) Spatial distribution regions of the elderly constructed by Voronoi diagram in Beijing

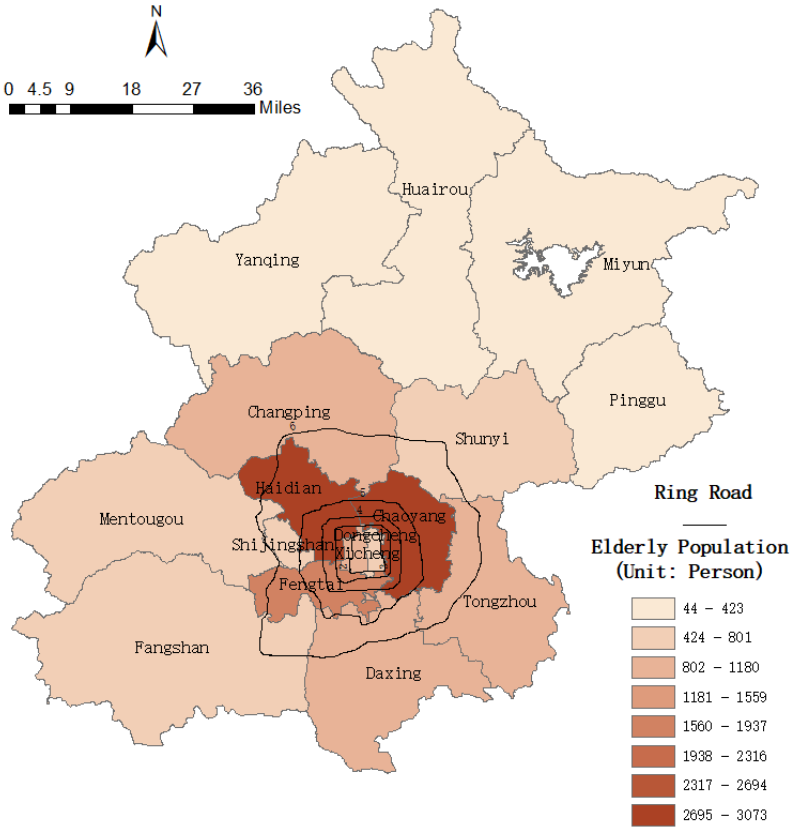
A comparison of two city partitioning methods for describing spatial distribution of the elderly

To demonstrate the rationality of constructing the spatial distribution of regions using the Voronoi diagram, we compare the spatial distribution regions using the Voronoi diagram with normal administrative regions.

In this application study, we present the spatial distribution of the elderly in the city of Beijing on two city partitioning methods, a) 16 administrative regions, b) 328 counties and c) a 147-polygon Voronoi diagram, in Figures 4.6, 4.7 and 4.8, respectively. The density level of the elderly population distribution in the different regions is indicated in different colors. The 16 administrative regions, which have commonly been used for spatial analysis in the past, are shown in Figure 4.6, where the yellow lines are Ring Roads. The color scale indicates the population density of the elderly. It can clearly be seen that the color is darker in the center than in remote regions. However, the area of each administrative region is too large for the elderly spatial distribution to be illustrated in detail.



(a) Spatial distribution of the elderly population according to government statistical data by the 16 administrative regions



(b) Spatial distribution of the elderly population according to detected result by the 16 administrative regions

Figure 4. 6 Spatial distribution of the elderly population by the 16 administrative regions

Figure 4.7 shows the spatial distribution of the elderly in 328 counties, with much more detailed information due to the smaller area of these divisions; the black lines are Ring-Roads. From the spatial distribution of the color scale, we can see that high-density regions are distributed slightly away from the center. Even though these small regions provide more detailed information, they are so numerous that the number of the elderly within each region is very small.

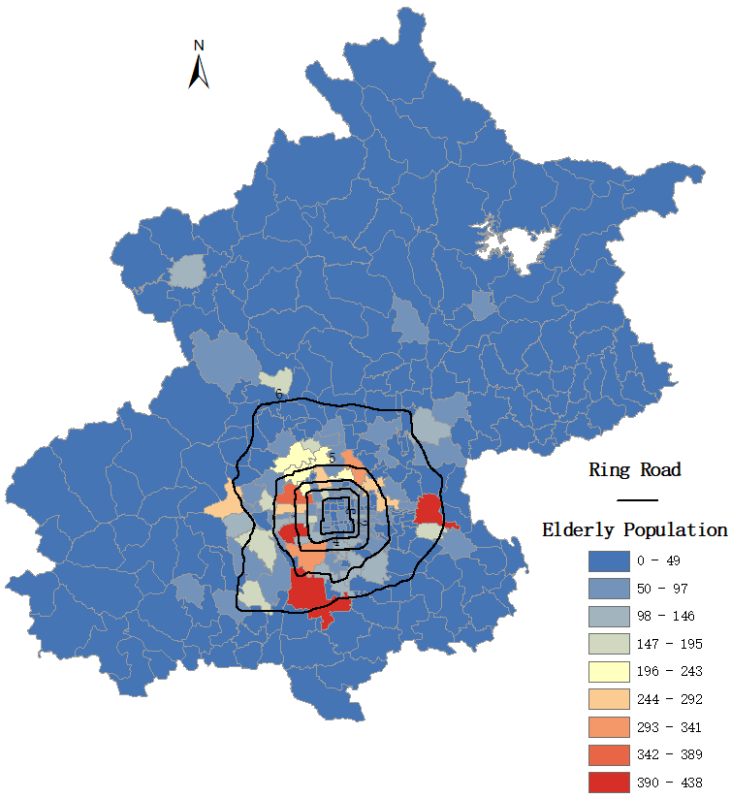


Figure 4. 7 Spatial distribution of the elderly population in 328 counties

Figure 4.8 shows the spatial distribution of the elderly population using the 147-polygon Voronoi diagram, where black lines are Ring Road. From the distribution of the colored regions, we can see that the regions with high population density are concentrated in the center of the city. Remote regions have lower population densities. The area of each region in the Voronoi diagram is smaller than the administrative region, and thus the elderly spatial distribution is illustrated in greater detail.

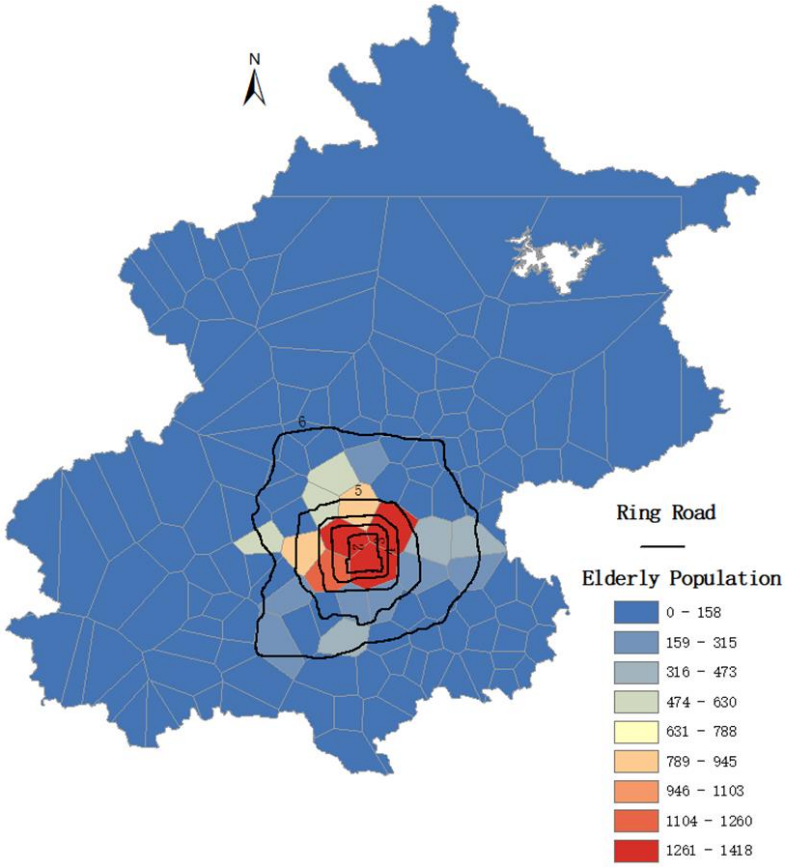


Figure 4. 8 Spatial distribution of the elderly population in a 147-polygon Voronoi diagram

To evaluate the quality of the spatial distribution of the elderly population described by the Voronoi diagram compared with that by administrative regions, Moran's I, the Z-score and the P value were used to calculate the spatial correlation for each of the spatial representations. To benchmark these results, we used data from white papers on the Development of Aging Service and Care System in Beijing from 2017 (Beijing Committee on Aging, 2018). Table 4.6 presents the spatial correlation results based on Moran's I for the three presentations.

Table 4. 6 Comparison of different three presentations on elderly distribution

	Moran' I	Z-score	P value	Distribution Mode
Figure 4.6 (a)	0.365	2.783	0.005	Clustered
Figure 4.6 (b)	0.017	0.545	0.585	Random
Figure 4.7	0.126	10.037	0.000	Clustered
Figure 4.8	0.467	15.840	0.000	Clustered

We can clearly see in the graphical visualization of the elderly population distribution has clustered characteristics. Most elderly residents live in the central regions of the city. The second row of Table 4.6 (a) also supports this feature of the distribution. Moran's I has a positive value of 0.365, which means that the spatial distribution of the elderly is positively correlated. The P value is 0.005, which is smaller than 0.01, and the Z-value is larger than 1.96. These statistics also indicate that the elderly distribution has the clustered characteristics. We regard these graphic and statistical findings based on official census data on spatial distribution of the elderly as a reference benchmark, and compare the other two following spatial partition methods based on smart card data with this benchmark.

The third row contains the calculated spatial correlations for the 16 administrative regions based on smart card data. The Moran's I value is only 0.017, which is very small, meaning that the population distribution does not have a significantly positive correlation. The P value is 0.585, which is larger than 0.1, and the Z-value is less than 1.65. These results conclude that the elderly population is more like a random rather than a clustered distribution, which contradicts with the benchmark finding above.

The distribution over the 328 counties has similar characteristics to the a 147-polygon Voronoi diagram, although the latter shows a more obviously clustered pattern than the former. Moran's I for the Voronoi diagram is calculated to be 0.467 and that for the 328 counties is 0.126, and the Z-values are 15.840 and 10.037, respectively. The spatial distribution of the elderly by county can be explored in detail because of the large number of county divisions. However, in some counties the detected elderly populations are too small to be useful for analysis. The Voronoi diagram enable consideration of the number of elderly commuters getting on and off at each bus stop, and as a result, it provides a more appropriate regional division of the city.

The 147-polygon Voronoi diagram based on smart card data shows an obviously clustered pattern, and this is consistent with the benchmark finding above. Moran's I for the Voronoi diagram is calculated to be 0.467, and the Z-values are 15.840, respectively. The Voronoi diagram enable

consideration of the number of elderly commuters getting on and off at each bus stop, and as a result, it provides a more appropriate regional division of the city for representing spatial distribution of the elderly.

In summary, we used official census data as the benchmark to evaluate the quality of the regional divisions by two different methods, administrative regions and Voronoi diagram regions, to illustrate the spatial distributions of the elderly population. Moran's I, the Z-score and the P value were computed as indicators to evaluate their qualities. First, according to the visualized and numerical results, the benchmark clearly shows the clustered pattern of the spatial distribution of the elderly population. The patterns indicate that elderly residents are mostly concentrated in the center of the city. The calculated results show that the 147-polygon Voronoi diagram provides the most appropriate spatial division of the elderly population distribution. Using this diagram, the elderly distribution was divided into regions with suitable sizes and numbers, unlike when using the 16 administrative regions, the results of which are inconsistent with the benchmark findings.

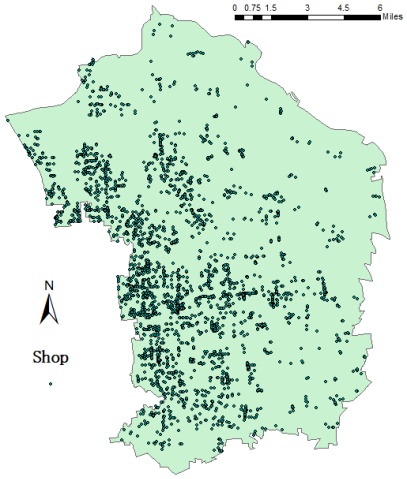
Comparing the three representations, we can conclude that the spatial distribution of the elderly population described by the 147-polygon Voronoi diagram is more appropriate and has the following two advantages: a) its spatial distribution of the elderly population is more representative, and b) it provides more detailed information on the distribution.

Another merit of the spatial distribution analysis of the elderly based on smart card data is up to date, say can be data of the day, or in last week, comparing with the census data for each administrative region which may be for last year or five years ago, for instance. Therefore, in the following sections the PoI-based elderly livability index analysis and connectivity between the elderly living regions will be based on the Voronoi diagram.

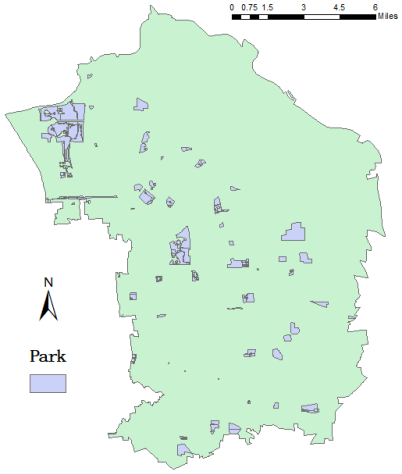
4.5.2 Explaining clustering distribution of the elderly by PoI-based elderly livability index analysis

PoI data

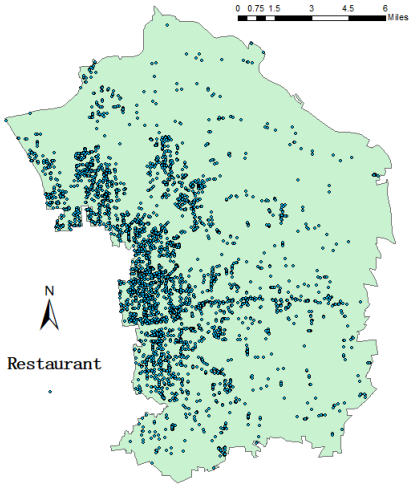
PoI data is used to represent the distribution of public facilities and calculate the elderly PoI-based elderly livability index of each region. We collected the more than 40 thousand data from website range of Beijing city includes five categories of PoI data namely shops, parks, restaurant, hospitals and bus stops. The PoI distribution of Chaoyang, one of the regions with the highest elderly population concentration in the city, as an example is shown in Figure 4.9 from (a) to (e), respectively.



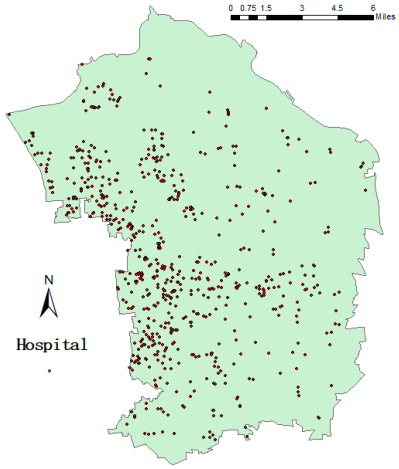
(a) Shop



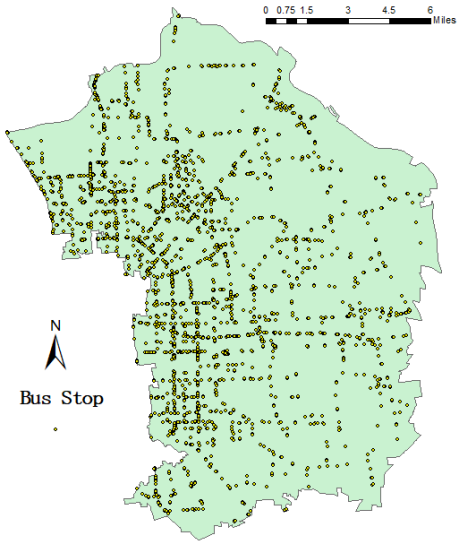
(b) Park



(c) Restaurant



(d) Hospital



(e) Bus Stop

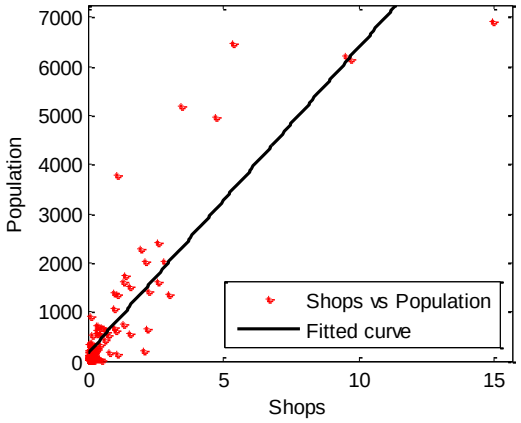
Figure 4. 9 PoI data of Chaoyang region

Correlation of the elderly distribution and PoI distribution

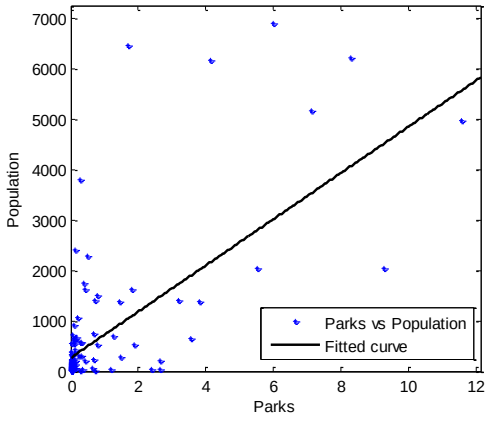
The proposed PoI-based elderly livability index model is used to analyze and explain the spatial distribution pattern of the elderly. PoI-based elderly livability is a complex issue and may be related to many factors, including social, economic, security and environmental factors, in addition to public facilities. In this study, we focus on public facilities because the quality of life in residential areas is strongly dependent on the density of public facilities. We also consider the issue of data availability for computing the index.

In the proposed the PoI-based elderly livability index, we identified five kinds of PoI on public facilities: shops, parks, restaurants, hospitals and bus stops. To characterize the PoI distribution of each region, we counted the total number of shops, restaurants and bus stops in each polygon of the Voronoi diagram. The total area of parks was assigned as the index. We assigned the weight for each hospital according to these calculated levels. The PoI distribution clearly shows a clustered pattern in which some regions have more PoIs than others, and most are located in the center of the city.

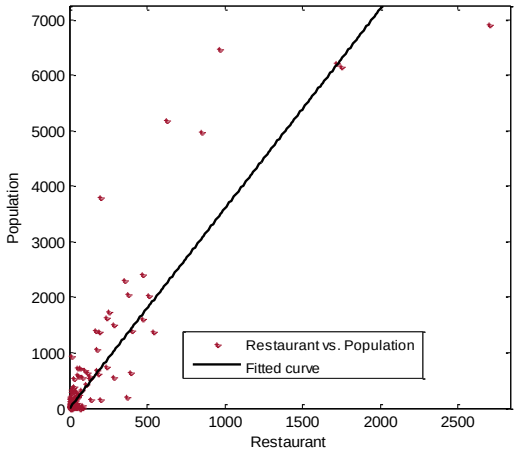
The Pearson correlation coefficients between the elderly population and each category of PoI were calculated. Figure 4.10 shows the results of curve fitting using each kind of PoI and the elderly population. Most points are distributed close to the fitted curves with $R^2 > 0.7$, except in Figure 4.10(b). This indicates that the distributions of PoIs and elderly population are positively, geographically correlated. We calculated the total area of parks as the indicator when processing the PoI data, instead of calculating the number of parks, as regions with large areas of parkland necessarily have smaller residential areas, which in turn leads to a low population. All in all, the PoI distribution was correlated with the population distribution. To evaluate the correlation, the Pearson coefficients of PoIs were calculated as shown in Table 4.7. From the table we can see that the relationship between PoI and population fits a linear equation very well.



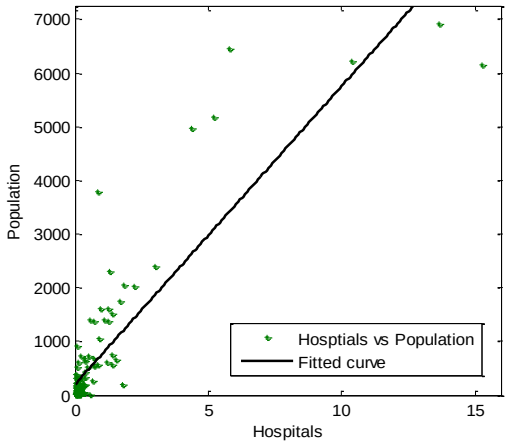
(a) Shops (R^2=0.78)



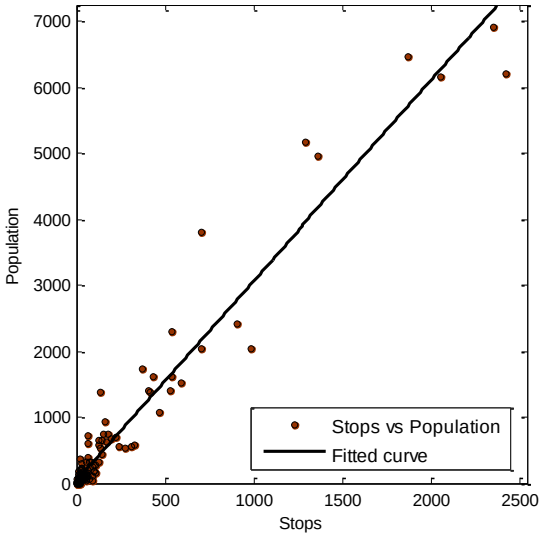
(b) Parks (R^2=0.41)



(c) Restaurants (R^2=0.78)



(d) Hospitals (R^2=0.76)



(e) Bus stops (R^2=0.95)

Figure 4. 10 Dependence of PoI on population

Table 4. 7 Coefficient of PoI and population

PoI	Pearson correlation coefficient
Shops	0.8925
Parks	0.6750
Restaurant	0.9001
Hospitals	0.8856
Bus stops	0.9741

The PoI-based elderly livability index distribution

The rationale for the selected five factors for the PoI-based elderly livability index in terms of PoI was verified by calculating the correlation of the elderly population and PoI in the region. The PoI-based elderly liability index of each region was then calculated according to equation above.

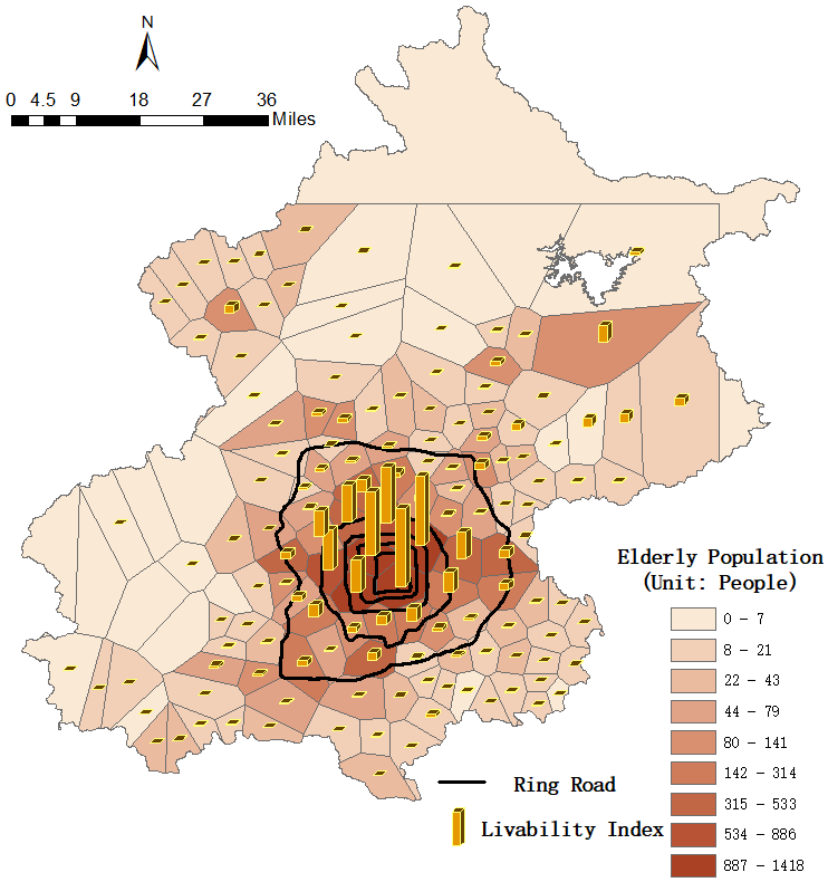


Figure 4. 11 Spatial distribution of elderly population and the PoI-based elderly livability index

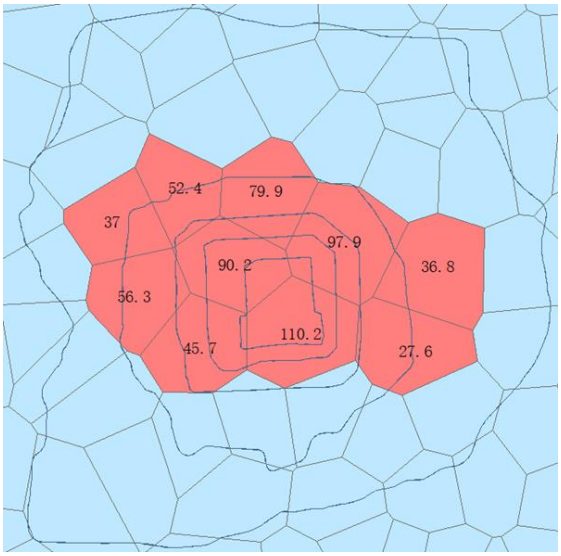


Figure 4. 12 Ten regions with higher PoI-based elderly livability index values

Figure 4.11 shows the spatial distribution of the elderly population together with the PoI-based elderly livability index. Yellow bars indicate the PoI-based elderly livability index of each region and the various red colors of the polygons in the Voronoi diagram represent the different population densities of the elderly. From the figure we can see that higher values of population and PoI-based elderly livability index are mainly concentrated at the center of the city, and lower values in the remote regions. This reflects the fact that the density of elderly residents is related with the availability of public facilities indicated by the PoI distribution, since the level of public facilities density are also high. This explains the fact that the elderly prefers to live in high-quality residential areas with comprehensive public facilities. Figure 4.12 shows the 10 regions with the highest PoI-based elderly livability values. These regions are likely to be the first choices when elderly residents choose where to live. It can be clearly seen that the northwest regions of the city center have the highest PoI-based elderly livability indices. These 10 regions span from north of the Fifth Ring Road to south of the Fourth Ring Road, and from east to west of the 5th Ring Road.

This spatial distribution of the elderly and PoI-based elderly livability index can help the government planning department to make urban development policy. As the distribution of public facilities is unbalanced, the elderly living in remote places without sufficient facilities must travel

long distances to obtain what they need. The planning department should therefore plan more public facilities to meet the needs of the elderly. The spatial distribution clearly shows the regions in need of development. Even though the center regions have sufficient facilities, their elderly population is large. It is better to enable the elderly to move to remote regions by developing public facilities there. If it is difficult for the elderly to move to other places because of poor physical health, more care centers, such as nursing homes, should be established.

In summary, the distribution of PoI-based elderly livability indices across Beijing calculated from the POI distribution is highly unbalanced, with the most livable areas located in the center of the city. This could easily lead to traffic congestion and high population density in particular regions, which may render these locations unsafe for public events that are popular with the elderly. However, they provide convenience for elderly residents by removing the need to travel long distances.

4.5.3 Analysis of connectivity between the elderly living regions by network analysis

Network construction and analysis

A network was constructed based on the centroids of the 147-polygon regions in the Voronoi diagram, and study on the connectivity between the regions with respect to the activities of elderly was conducted accordingly. The coordinates of the centroids for each polygon were regarded as network nodes. The locations of bus stops within each region were aggregated to the centroids. If any elderly citizen took a bus from one region to another, the linkage between the two centroids was regarded as a network edge. A spatial connection network of elderly commuters was constructed and is shown in Figure 4.13.

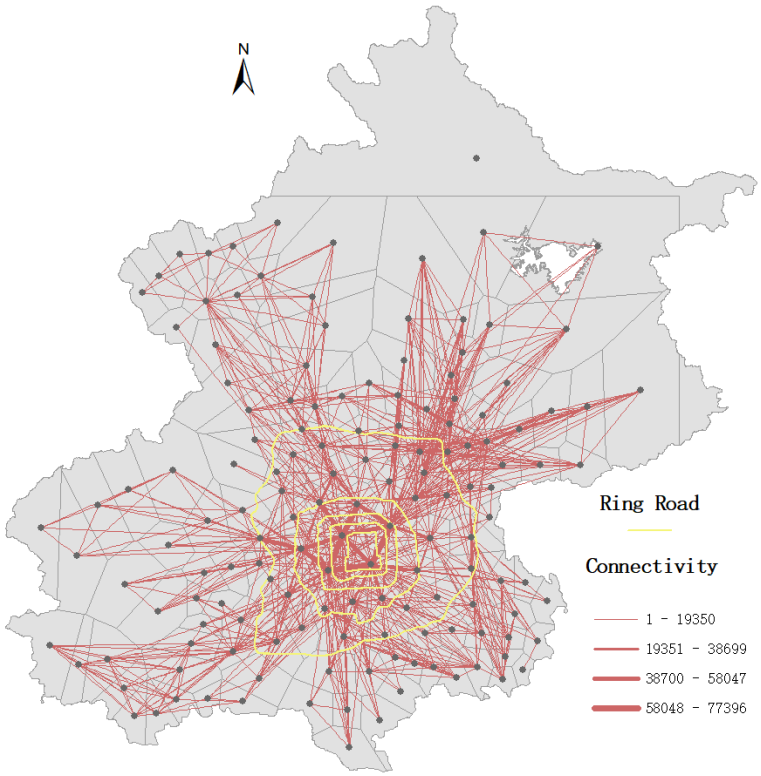


Figure 4. 13 Connectivity between the elderly living regions

In Figure 4.13, the grey regions are the 147-polygon region of the Voronoi diagram, the yellow lines are Ring-Roads, the light grey dots are the centroids of the polygons in the Voronoi diagram and the red color lines are the connectivity between pairs of polygon regions related by the activities of the elderly. If there was no connection between any two regions, the two nodes are not connected by any line. The thicker the red lines, the greater the connectivity between the two nodes. We can observe that most of the strong connections are within the 4th Ring Road of the city. The connections among remote regions beyond the 5th Ring Road are weaker. We can also see that pairs of regions with strong connections are mostly very close to each other. With increasing distance between two regions, their connections gradually weaken.

To explore the network characteristics on connectivity within polygon regions in the Voronoi diagram, we calculated the network properties from global to local scales. Table 4.8 illustrates the overall properties of the Voronoi diagram, and Table 4.9 presents the connectivity properties of each polygon region in the diagram.

From a global point of view, the whole city is divided into 147 regions, the same as the number of nodes. In the network, only 972 edges are detected, which is lower than the total possible connectivity, 10,731. There are three reasons for this. The first is that not all pairs of regions are directly connected by bus lines. The second is that some remote regions can be accessed by other public transportation modes that are not recorded in our smart card data source. Third, terrain could be an influential factor, as features such as mountains and rivers form natural obstacles to setting up a bus line.

Table 4. 8 Overall network properties of Voronoi diagram

Voronoi diagram	
Nodes	147
Edges	972
Strength	1,485,797
Loops	1,679,231
Average degree	25
Average strength	20,633
Graph density	0.09

Table 4. 9 Network properties of each region in Voronoi diagram

Centroid longitude	Centroid latitude	In degree	Out degree	In strength	Out strength	Loop strength
1,027,561	2,769,439	48	50	159,017	145,373	40,544
1,020,676	2,776,289	38	42	188,443	178,019	36,852
1,031,977	2,778,604	48	50	124,913	130,885	34,369
1,017,376	2,768,020	41	40	129,312	129,409	31,826
1,011,181	2,773,155	34	36	94,593	102,497	31,794
1,024,287	2,783,708	31	30	110,996	108,020	30,759
1041347	2,775,720	28	28	32,894	35,378	27,897
1,050,981	2,775,882	21	22	23,182	23,122	27,576
…	…	…	…	…	…	…

Two kinds of connection patterns are important for understanding the elderly connectivity: connections within each region and the connections between regions. In this network analysis, the total number of connections between each pair of regions is taken to represent the strength of the connectivity between regions, while loops represent the total number of connections within each region. Thus, the strength value is 1,485,797 while the loop value is slightly greater at 1,679,231. Based on the result we can see, from a global view, even though most of the traveling by elderly happens within regions (46.9%), this form of travel is not obviously different from travel between regions (53.1%). From the local view, the strength and loop number of each region displays various patterns. For example, the in-strength and out-strength of regions are larger than the loop values. At the bottom of the table, however, each region has a similar strength value to its loop value. This indicates that in some regions most of the travel by the elderly is between regions rather than within those regions. Therefore, from a global view the connectivity of the Voronoi diagram is unbalanced.

We can see from Table 4.8 that the average degree is 25, which means that each region is on average connected with 25 other regions. This means that the elderly lives in one region travel a lot to 25 other regions. We can observe from Table 4.9 that few regions have large values of degree. Each region has approximately 20,633 connections on average, according to the average strength value. The graph density is approximately 0.09, which means it is a sparse graph with only a few edges. No region is connected with all other regions. Compared with remote regions, those located within the Fourth Ring Road are connected with a larger number of other regions, which is probably a consequence of urban planning and development. That is, Beijing is an ancient city with a long history, and with the ongoing development of urbanization, public facilities and commercial centers are blossoming in the Forbidden City at the heart of Beijing and gradually spreading outward. This leads to unbalanced urban development with respect to the areas where the elderly take part in activities.

The connectivity between the elderly's living regions describes their travel frequency between regions. Many public transports such as subways and buses have priority seating for people with special needs. However, the traffic flow is so huge in this large city that it may be not enough for

the elderly. Considering their physical health and their low travel frequency, customized shuttle buses would be a good solution for the elderly. The transportation department should thus develop a special public transportation line and schedule in line with the connectivity results to make the elderly's travel more convenient.

Relationship between PoI-based elderly livability index and connectivity level by the quasi-gravity model

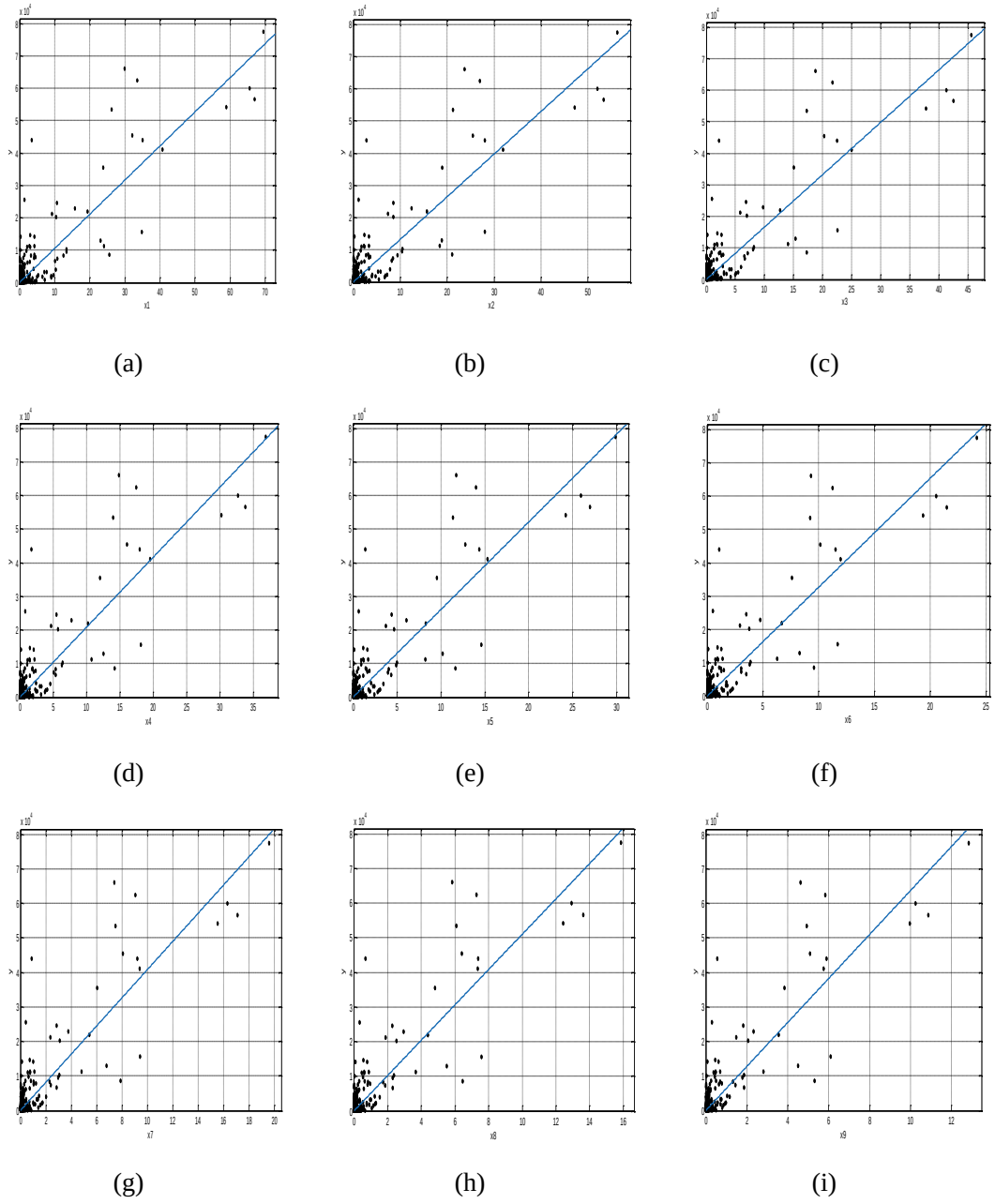
In this chapter, the proposed quasi-gravity model was used to quantify the relationship between the PoI-based elderly livability index and connectivity between pairs of regions. In the chapter, three factors were imported into the model. The strength of the network is regarded as an indicator of mobility. The PoI-based elderly livability index corresponds to the PoI indicator. Euclidean distance was used to calculate the distance between two nodes, but the square of the distance may not be the optimal order for selection.

Table 4. 10 Various parameters' results of developed model

Figure 4.14	G	γ	R^2
(a)	1055	2.2	0.7660
(b)	1324	2.3	0.7685
(c)	1662	2.4	0.7704
(d)	2083	2.5	0.7718
(e)	2609	2.6	0.7726
(f)	3265	2.7	0.7730
(g)	4084	2.8	0.7729
(h)	5104	2.9	0.7724
(i)	6375	3.0	0.7715

To find the optimal parameters of this model, curve fitting was adopted to fit the values of γ and G. Using estimated values of γ ranging from 2.2 to 3.0 as given in Table 4.10, the corresponding values of G were computed. Figure 4.14 shows the fitted curves of the model, where the X axis is

L_Index_i L_Index_j/D^γ, and the Y axis is the strength between two regions S_{ij}, and the blue lines are fitted curves. According to the table, with increasing γ, the goodness-of-fit R^2 first reaches a peak value of 0.7730, then decreases. We selected the fitness value of γ as 2.7 and the value of G as equal to 3265, giving R^2 its maximum value of 0.7730.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4. 14 Fitted curves of quits-gravity model

Finally, we selected the optimal model, which is represented as Equation (4.15):

$$S_{ij} = 3265 * \frac{L\_Index_i \, L\_Index_j}{D^{2.7}} \quad (i \neq j, i = 0,1,2,\dots, j = 0,1,2,\dots) \tag{4.15}$$

In summary, the mobility of the elderly in Beijing can be quantified with a quasi-gravity model above based on PoI data and the distance between regions. We can conclude from the model that distance is not a major factor affecting the purposes of travel for elderly. Thus, if one region has a large PoI-based elderly livability index value, while another has a small value, the strength between the two regions is unlikely to be affected even if the distance between them is large. This motivates these elderly people to travel great distances to access high-quality public facility services including shops for necessities.

With the development of public transportation systems, convenient and efficient buses and subways can make long-distance travel less challenging for the elderly, even in a large city. The public facility services available in any single residential area may not satisfy the increasing material and cultural demands of elderly for instance. However, long-distance travel for the elderly may be routine in the cities of the future.

4.6 Conclusion

This chapter presented a framework including a series of new data-driven methods for studying the spatial distribution of the elderly population to address the issue of the population ageing, a trend and challenge for cities worldwide now and future. Specifically, this framework systematically answers the following three questions: a) Where do they live? b) Why the elderly distribution like this? c) Where do they go frequently?

To answer the question of the spatial distribution pattern of the elderly, three methods are developed: (a) a method for identifying the home locations of the elderly based on bus stop locations, (b) Voronoi diagram-based method for partitioning a city into regions that reflects the accuracy of the elderly living, that is not based on administrative regions and (c) an enhanced method to cluster bus stops based on the elderly's flow at stops and to create a Voronoi diagram based on the cluster

centers.

To answer the question of why the elderly distribution like this, an PoI-based elderly livability index model was proposed to measure and explain the important factors for elderly citizens when choosing their home location, and to rationalize the spatial distribution of the elderly population. In this chapter, the public facility aspect, and restaurants, parks, hospitals, shops and bus stops were identified as the key factors for the PoI-based elderly livability index model. The model can compute an indicator of the level of PoI-based elderly livability of each region for the elderly.

To answer the question of where the elderly go frequently, spatial connectivity was used to analyze the elderly travel behavior between two regions, including, for example, from home locations to hospitals. Connectivity links the region to which elderly citizens travel from their living region and where they undertake activities. These are the primary regions to which the elderly travel. Furthermore, a quasi-gravity model was developed and confirmed to be valid for the relationship between the spatial connectivity between any two regions in terms of network strength and the PoI-based elderly livability index of the different regions concerned.

As an application of the proposed framework and methods developed in this chapter, Beijing was identified as the case study area, and smart card data, PoI data and bus stop location data were used to analyze the spatial distribution of elderly citizens in the city. Three findings arose from this chapter: a) the spatial distribution of the elderly shows clear clustering characteristics; b) the spatial distribution of the elderly has a strong relationship with that of public service facilities, such as restaurants and hospitals; and c) the connectivity of each pair of regions is related to the distribution of public facilities in the connected regions.

The significance of this chapter lies in its development of the framework and series of methods for comprehensively understanding the spatial distribution patterns of the elderly in a city, which is essential for urban planning, management and services given the population ageing trend worldwide.

Chapter 5 Analytics of temporal characteristics of the elderly mobility

Chapter 4 described the spatial distribution characteristics of the elderly. In this chapter, the temporal characteristics of the elderly mobility is analyzed by statistical description method and distribution function of curve fitting. This chapter investigated the current mobility behavior of the elderly using smart card data. Their behavior was analyzed via four spatiotemporal features: departure and arrival time, travel distance, duration, and frequency. The analytics for the elderly was also compared with the adults' group. Beijing, a city with very high life expectancy, was selected for the study. Two methods were adopted for the analytics: (a) a quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent graphical presentation of such behavior, and (b) the discovery of the distribution function of the travel characteristics, both by function curve fitting and by testing the goodness of fit of the identified distribution functions.

5.1 Introduction

Many countries around the world face the challenge of an ageing population. A full understanding of the mobility behavior of the elderly is a necessary and urgent requirement to enable cities to plan for and cope with this trend. It is expected that this trend will have significant social, economic, and cultural global effects. To manage the smooth operation of a megacity, it is necessary to carefully consider the effects of one of its most influential factors, the increasing ageing population. Specifically, it is necessary to understand that the major aspects of the elderly's mobility behavior are likely to be associated with specific health difficulties and social problems related to age. Thus, to make living less of a burden, both for the elderly themselves and for others, effective transport planning and traffic management would not only benefit society as a whole but would also produce a general feeling of independence, which itself breeds harmonious feelings and even the possibility of reducing the forms of crime produced by frustration.

This chapter investigated the current mobility behavior of the elderly using smart card data. Figure 5.1 shows the logical flow of this chapter. The elderly behavior was analyzed via four spatiotemporal features: departure and arrival time, travel distance, duration, and frequency. The analytics for the elderly was also compared with the adults' group. Beijing, a megacity with very high life expectancy, was selected for the thesis. Two methods were adopted for the analytics: (a) a quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent graphical presentation of such behavior, and (b) the discovery of the distribution function of the travel characteristics, both by function curve fitting and by testing the goodness of fit of the identified distribution functions. The significance of this chapter is its contribution to future smart city planning, management, and services for the elderly in a megacity, based on a close observation of the daily travel behavior of the elderly.

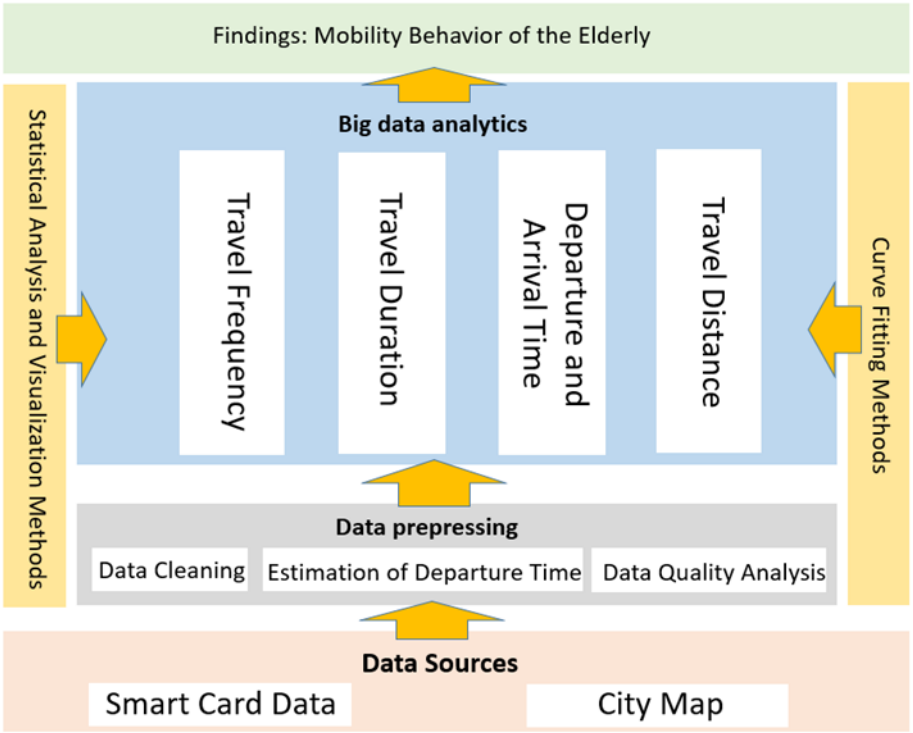


Figure 5. 1 Logical flow of the elderly behavior analysis

5.2 Methodologies

5.2.1Distribution functions

In this chapter, as described earlier, the mobility characteristics of the elderly are studied in terms of travel distance, travel time, travel duration and travel frequency. Two approaches will be used: a) statistical analysis and graphical visualization; and b) function curve fitting. Statistical analysis is used to describe the basic distribution of the identified travel characteristics. This is an intuitive method and the distribution can be further interpreted via graphical visualization. Function curve fitting is a method to quantify the distribution of travel distribution by mathematical models.

It is essential to identify an appropriate function to fit the distribution of the mobility characteristics from the available functions that could statistically fit the distribution of the travel characteristics. These include exponential, Gaussian, and other types. These functions are tested individually to enable the identification of the function best fitted to determine the distribution of the mobility characteristics. The travel distance and travel duration of the elderly are included, based on data from the smart card. What follows are the candidate functions to enable identification of the later function curve-fitting analytics.

The exponential distribution is as follows:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{5.1}$$

where λ denotes the rate parameter, *x* stands for the travel distance, and f(x) stands for the statistical distribution given distance x.

The Gaussian distribution function is presented as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \alpha \exp\left(-\frac{x-\beta}{\gamma}\right)^2 \tag{5.2}$$

where μ and σ denote the mean and standard deviation of the data set, respectively. The right-hand side is a simple method to describe the function.

The power law distribution is defined as

$$f(x) = x^{-\alpha} \tag{5.3}$$

where α is the exponent.

The gamma distribution is presented as

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0 \tag{5.4}$$

in which α denotes shape parameter, β denotes scale parameter, $\Gamma(x)$ denotes gamma function.

The Weibull distribution is expressed as

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(x/\lambda\right)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{5.5}$$

where λ is scale parameter, k is shape parameter.

The mobility behavior of the elderly is analyzed based on the smart card data from the elderly and adults. Their respective mobility behavior, the uniqueness of the elderly, and their similarity with adults in terms of mobility characteristics will thus be identified. In this chapter, the data set used

for analysis is the Beijing smart card data set for April 12, 2017, after cleaning and preprocessing.

5.3 Mobility characteristics of the elderly

5.3.1 Travel distance

Definition of travel distance

In this study, the distance of the trip along the real routes between the departure and arrival stops is adopted as the travel distance. Specifically, this is defined as the distance between the departure or original bus stop (O) and the arrival or destination bus stop (D) for a person using that bus trip along a city bus route network. The travel distance of each O-D pair is calculated, and the distance computed by this method reflects the actual distance a bus travels along a given bus route network.

Statistical description of travel distance

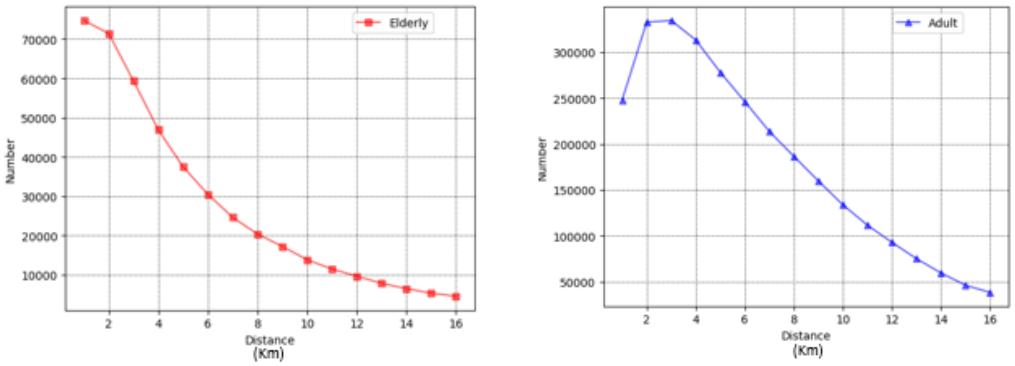
Based on the smart data for 441,560 trips by the elderly and 2,869,493 trips by adults, the statistical descriptions of the travel distance characteristics of the two groups were obtained and are illustrated in Figure 5.2 Figure 5.2(a) shows the statistical description of the travel distance for the elderly, and Figure 5.2(b) shows the travel distance for adults.

The travel distance characteristics of the elderly is summarized in the first row of Table 5.1 and in Figure 5.2(a). The travel distance of the elderly ranged from 1 to 16 km. The mean and median values for the travel distance were 4.9 km and 4 km, respectively (SD, 3.7 km). For this study, the median is seen to be more reliable and is recommended.

To compare the travel distance results of the elderly and those of the adults, the adults' travel distance characteristics were also calculated of 2,869,493 trip samples. The statistical results are listed in the second row of Table 5.1, and a graphical description is given in Figure 5.2(b). The travel distance for adults ranged from 1 km to 16 km. The mean and median values for the travel distance were 6 km and 5 km, respectively. A standard deviation is 3.8 km.

Figure 5.2(a) shows the trend of travel distance for the elderly. Of more than 400,000 trips by the elderly, some 70,000 trips exceeded 1 km and about 15,000 trips exceeded 2 km; a sharp decrease was seen in the number of trips in their travel distance from 2 km to 12 km; 1000 trips went 12 km; and trips longer than 2 km decreased from 400 to very small with the maximum travel distance of 16 km.

Figure 5.2(b) shows the trend of travel distance for the adults. The peak of adults' travel distance was 3 km. The number of trips increased from 1 km to 3 km, and a sharp decrease in the number of adults' trips was seen from 3 km to 16 km. Among more than 2,869,000 adult trips, 250,000 adults' trips were 1 km; about 370,000 adults' trips were 3 km (the peak); about 25,000 adults' trips were 6 km; 18,000 adult trips were 8 km; and the adults' trips of more than 15 km decreased from 4 k to very few with the maximum travel distance of 16 km.



(a) The Elderly (b) The Adults

Figure 5. 2 Statistical description of travel distance based on smart card data

Table 5. 1 Statistical parameters of travel distance data (Unit for distance: km)

	Number	Minimum	Maximum	Median	Mean	STD
Elderly	441,560	1	16	4	4.9	3.7
Adults	2,869,493	1	16	5	6	3.8

In summary, most the elderly's travels by bus are within 1 km, and the median distance is 4 km,

which is shorter than that of the adults (5 km). There is sharp decrease in the number of the elderly from 2km to 12km. The number of the elderly travelling more than 12 km is small.

Fitted function for travel distance

The function curve fitting method was used to analytically discover the regularity of travel distances for the elderly. By testing all possible distribution functions, the exponential function was identified as the best fitting model for the travel distance distribution of the elderly and distance thus follows an exponential distribution.

Table 5.2 shows the parameters and goodness of fit. The SSE is the sum of squares due to error. RMSE is the root mean squared error. R square is the coefficient of determination. In this case, for the distribution, the goodness of fit R square value reaches 0.9905, which indicates that the exponential distribution function fit the travel distance characteristics of the elderly well, in accordance with the smart card data.

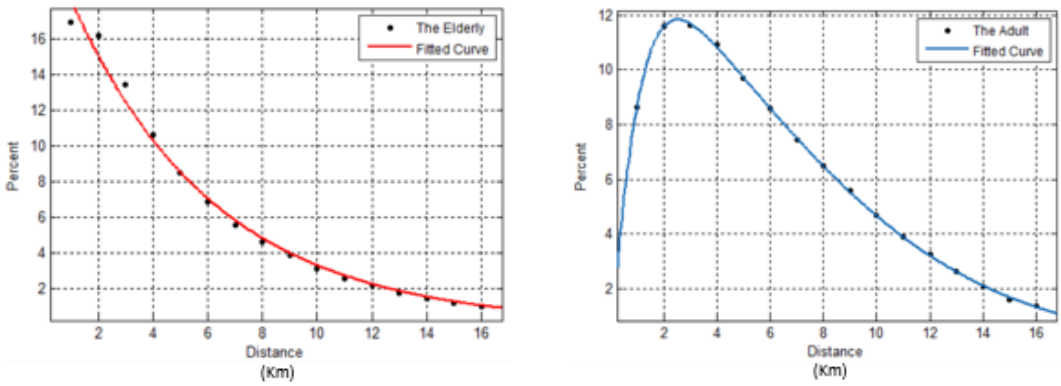
The regularity of the travel distance of the adults was also studied at the same time by the function fitting method. Again, testing of the possible distribution functions showed that the Gaussian function was the best fit model for the travel distance distribution of adults and thus follows the Gaussian distribution (R square, 0.9999).

Figure 5.3 shows the probability distribution and fitted curve. Figure 5.3(a) gives the fitted curve that represents the data of the elderly, and Figure 5.3(b) gives the fitted curve for the adults.

In Figure 5.3(a), the black dots represent the percent of the elderly who travel a given distance, and the red line represents the fitted function between the percent and the travel distance. The red line in Figure 5.3(a), which is an exponential function, fits the black dots well, which indicates that the exponential function graphically assesses the travel distance distribution of the elderly. The trend of the red curve in Figure 5.3(a) clearly shows a sharp decrease in the number of the elderly travelling as the trip distance increases.

In Figure 5.3(b), the black dots represent the percent of the adults who travel a given distance, and the blue line represents the fitted function between the percent and the travel distance. The blue line in Figure 5.3(b), which is a Gaussian function, fits the black dots well. This visualization result demonstrates that the Gaussian function can serve as a good graphic representation of the travel distance distribution.

Based on the fitted distribution functions with R square values and the graphic visualization results above, the following can be concluded: (a) the travel distance distribution of the elderly follows an exponential function, and (b) the travel distance distribution of the adults follows a Gaussian function.



(a) The Elderly (b) The Adults

Figure 5. 3 Fitted curves of travel distance

Table 5. 2 Fitted model of travel distance and goodness of fit

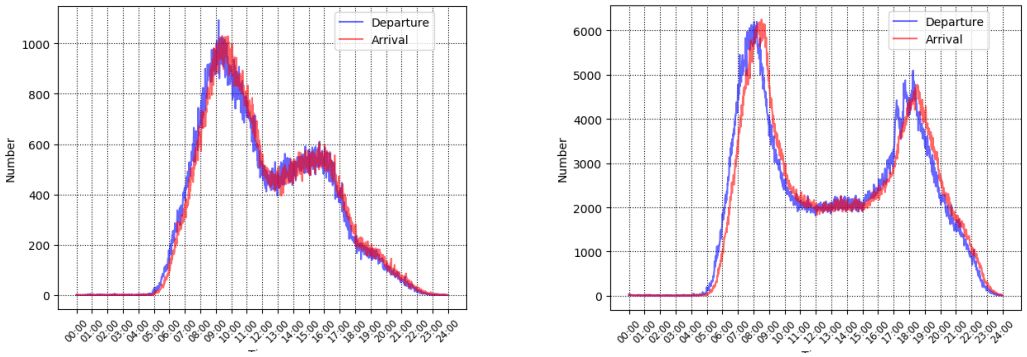
	Elderly	Adults
Model	$f(x) = 22.04 * e^{-0.1893*x}$	$f(x) = -81.76 * e^{(-(x+3.727/2.888)^2)} + 17.85 * e^{(-(x+5.247/13.16)^2)}$
Goodness of fit		
SSE	4.123	0.02194
R square	0.9905	0.9999
RMSE	0.5427	0.04684

5.3.2 Departure and arrival times

Temporal characteristics

Mobility characteristics have a strong relationship with the time attribute. Temporal characteristics are essential measurements for human mobility behavior. In this chapter, we adopt three travel characteristics related temporal attributes—the departure time, the arrival time, and the travel duration—to analyze the travel behaviors of the elderly and to compare them with those of the adults group. First, the departure time and arrival time are compared by visualizing their distributions. Second, the statistical description of travel duration is analyzed regarding the mobility characteristics, and a fitted curve function is generated and tested to describe the travel duration regularity.

Statistical description of departure and arrival times



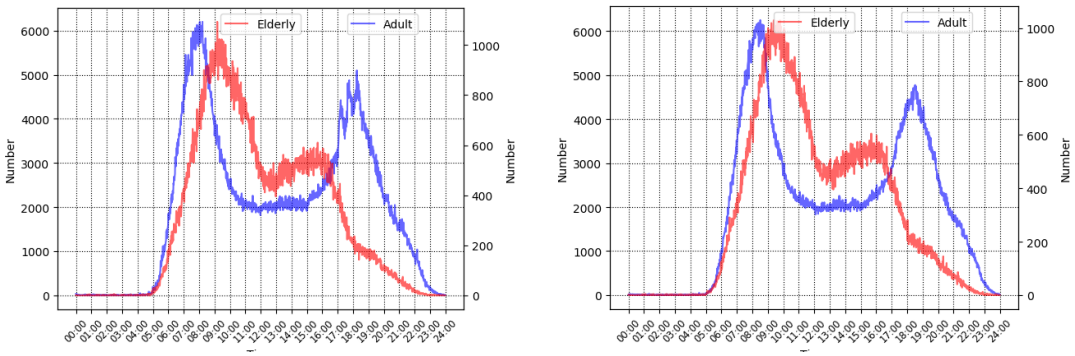
(a) The Elderly (b) The Adults

Figure 5. 4 Departure and arrival time where blue line represents departure time and red line represents arrival time

Figure 5.4(a) is a statistical description of the departure and arrival times of the elderly. The blue line represents the departure time, and the red line represents the arrival time. At first glance, the trends of the lines are very similar. Both have a clear peak from 9:00 to 10:00 AM. This period is the morning rush hour for the elderly's travel. The number of trips then decreases to a low level until 1:00 PM. From that time, there is a slight increase from 432 trips to 593 trips until 4:00 PM, followed

by a gradual decrease over time to zero.

The departure time and arrival time distributions of adults, however, differ from those of the elderly. Figure 5.4(b) describes the adults' departure time and arrival time distributions. Two clear rush hours can be seen, one in the morning and one in the afternoon. For the morning rush hour, the peak arrival time (around 8:35 AM) is a little later, but the period experiences a higher number of trips than the peak departure period (around 8:00 AM). The distribution curve shows that the duration of the rush hour in the morning is narrower than that in the afternoon, which indicates that many more people both alight from and remain on public transport for short durations, which can cause unpleasant crowding. In the afternoon, the departure time has three small peaks, but the arrival time has only one peak. Thus, it can be interpreted that even though people take public transport at slightly different times, they arrive at a similar time.



(a) Departure Time (b) Arrival Time

Figure 5. 5 A comparison between the elderly (in red) and the adults (in blue)

We now compare the distribution of the departure time between the elderly and the adults, as shown in Figure 5.5(a). The distribution of the number of people who used public transportation in each minute from 0:00 to 24:00 is presented in this Figure. The blue line represents the distribution for adults referring to the left Y axis, and the red line represents the distribution for the elderly referring to the right Y axis.

For the adults, based on Figure 5.5(a), two time periods are noted for the main departure peaks: 7:00

to 9:00 AM and 5:00 to 7:00 PM. Specifically, the morning peak begins at 8:00 AM. Three small peaks occur during the afternoon, namely at 5:13 PM, 5:48 PM, and 6:18 PM. This illustrates the habit of adults going to work in the morning within a relatively concentrated period. However, in the afternoon, to avoid rush hour traffic congestion, adults tend to choose public transportation earlier or later than the afternoon peak when possible. This could explain the three small peaks in the afternoon. The number of trips for the morning peak is 6200, and those for the afternoon peaks are 4186, 4876, and 5198, respectively.

For the elderly, only one clear peak is seen in the morning, at around 9:00 AM. The morning peak for the elderly is different from that of the adults in terms of time periods, but the time period overlaps from 8:00 to 9:00 AM. In the afternoon, a small increase in trip trends by the elderly occurs from 1:00 to 4:00 PM, before the number quickly decreases to a low level.

Table 5.3 clearly shows that the elderly travel more frequently in the morning than in the afternoon (56.07% vs 43.93%), whereas the reverse is true for the adults (45.34% vs 54.66%).

Table 5. 3 Total number of departures for the elderly and adults in the morning and afternoon

	Number	**Morning**	**Percentage**	**Afternoon**	**Percentage**	**Difference**
Elderly	441,560	247,573	56.07%	193,987	43.93%	53,586
Adults	2,869,493	1,301,049	45.34%	1,568,444	54.66%	267,395

Table 5.4 shows that 80% of the elderly complete their trip 3 hours before the adults; 90% of the elderly have an earlier arrival of 2 hours and 36 minutes; 95% of the elderly have an earlier arrival of 2 hours and 14 minutes, and 99% of the elderly have an earlier arrival of 1 hour and 31 minutes.

This can be interpreted as the elderly completing their daily activities earlier than the adults to get home earlier at night, whereas adults finish their activity late due to different job duties.

Table 5. 4 Time of completing the arrive trip for the elderly and adults

	80%	**90%**	**95%**	**99%**
Elderly	15:46	17:17	18:42	20:50
Adults	18:36	19:53	20:56	22:19
Difference	3 hours	2 hours 36 minutes	2 hours 14 minutes	1 hours 31 minutes
Average	-	-	-	**2 hours 20 minutes**

In summary, we can conclude that: (a) there is one peak for the elderly departures at 9:00 AM and no clear peak in the afternoon, but a relatively high number of departures between 5:13 and 6:18 PM; (b) the morning peak of departures for the elderly (9:00 AM) is 1 hour later than that for the adults (8:00 AM); (c) the elderly have more travel activities in the morning, whereas the adults have more travel activities in the afternoon; and (d) the elderly complete their travel activities earlier than the adults by approximately 2 hours.

Spatial distribution of bus stops at peak hour

To understand the spatial distribution of the bus stops used by the elderly and the adults during the morning peak hours, we present the bus stops used during the period by the two cohorts and visually analyze their overlap during the peak hours.

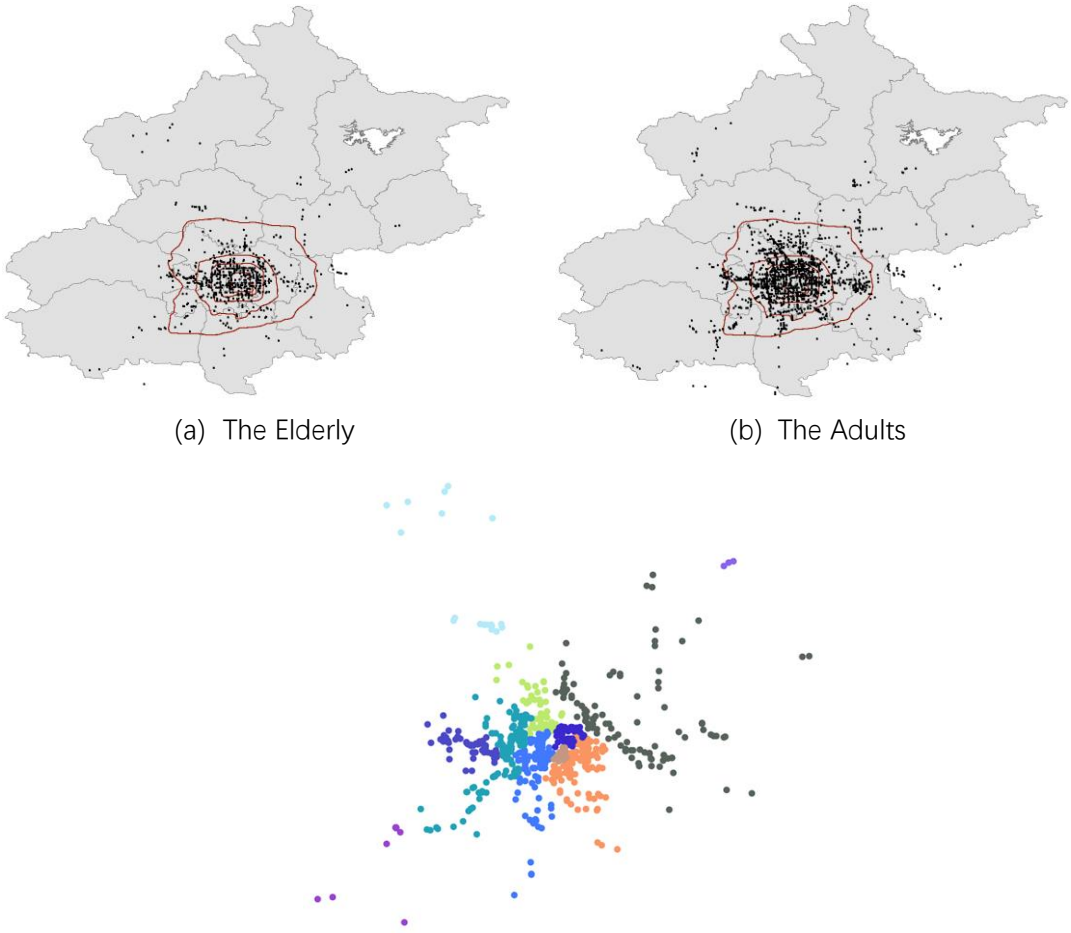


Figure 5. 6 Spatial distribution of departure stops during morning peak hour

Figure 5.6 shows the spatial distribution of departure stops during the morning peak hour for the elderly and the adults. Figure 5.6(a) shows the elderly's departure stops distribution at 9 AM, and Figure 5.6(b) shows the adult's departure stops distribution at 8 AM. Figure 5.6 (c) shows the cluster results of the elderly's departure stop by using the integrated clustering method. The total amount of adults' smart card data is greater than that of the elderly, so the density distribution of the adults' departure stops is thus higher than that of the elderly. Figure 5.6(a) shows that the stops are mainly concentrated within the 4th ring road, and the density is higher than that beyond the 4th ring road. The stops present a uniform distribution characteristic. Outside the 4th ring road, even though some stops show a clustering distribution characteristic in some places, the volume is small. Most stops are dispersed in these areas. Figure 5.6(b) shows that a greater density of the stop distribution, which is true not only for the area within the 4th ring road, but also for the area between the 4th and 5th ring

roads. Beyond the 5th ring road, there are several clear clusters with a large number of stops. Stops are even located in remote areas, but in much smaller numbers than in the central areas.

(a) The Elderly (b) The Adults

Figure 5. 7 Spatial distribution of the departure stops at 8:00 am in the morning

Figure 5.7 indicates a strong spatial location overlap of departure bus stops used by the elderly and the adults during the adults' morning peak (8:00 AM). This is especially true for the area between the 2nd and 4th ring roads, and the situation is even worse for the northwest area.

(a) The Elderly (b) The Adults

Figure 5. 8 Spatial distribution of departure stops at 9:00 am in the morning

Figure 5.8 shows the spatial distribution of departure stops at the elderly's peak hour (9:00 AM). There is no doubt that Figure 5.8(a) shows more used stops and a greater density of stops than Figure 5.7(a), which means that more elderly used these bus stops at 9:00 AM than at 8:00 AM. As for their spatial location, most of the used bus locations are still concentrated within the 4th ring road. In the

area between the 4th and 5th ring roads, the stop distribution still focuses on several clusters. For the remote regions outside the 5th ring road, the number of stops is much smaller. A comparison of Figure 5.8(a) and 5.8(b) reveals that the total number of the bus stops used by the elderly in Figure 5.8(a) is still less than that used by the adults in Figure 5.8(b). At 9:00 AM, a high number of bus stops is used by both cohorts within the 4th ring road. This indicates that both cohorts use bus services within the area of the 4th ring road at 9:00 AM, which could be one reason that traffic congestion is very high in the area during the morning peak. For the area outside the 4th ring road, the high density stops are distributed in several clusters, but the overall number is much lower than the area within the 4th ring road.

5.3.3 Travel duration

Definition of travel duration

As the temporal characteristic of mobility, the travel duration is an important measurement for analysis. The travel duration is defined as the period between a person boarding a bus and alighting from it. As mentioned in the data cleaning section above, it is assumed that the travel duration is less than 1 hour. The time is obviously based on the possible travel distance and traffic congestion conditions.

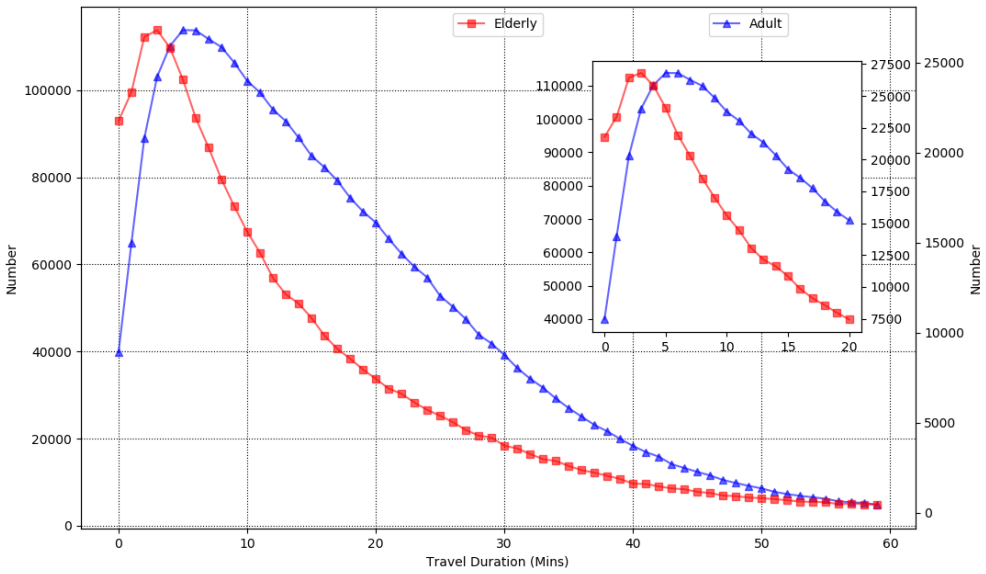


Figure 5. 9 A comparison of travel duration distribution based on smart card data: between the elderly (in red) and the adults (in blue)

Table 5. 5 Statistical parameters of travel duration (Unit for time: minutes)

	Number	Minimum	Maximum	Median	Mean	STD
Elderly	441,560	1	60	10	13.87	11.82
Adults	2,869,493	1	60	15	18.23	12.63

Figure 5.9 shows the statistical description of the travel durations of the elderly and the adults. The red line represents the elderly, and the blue line represents the adults. Figure 5.9 shows that the peak for the elderly is 4 minutes and that for adults is 6 minutes. This means that most the elderly stay on buses for 4 minutes and most adults stay for 6 minutes. Table 5.5 shows that the median travel duration for the elderly on buses is 10 minutes and that for adults is 15 minutes. The mean is higher because many people experience a longer travel duration. The reason for the shorter travel duration of the elderly is possibly due to health conditions or a dislike of compressed travel situations, whereas adults are time-controlled by their work environment.

From the observations and analysis above it can be concluded that (a) the travel duration of the

adults is double that of the elderly; and (b) most elderly stay on the bus for 4 minutes and most adults stay for 6 minutes. The median travel duration on public transport vehicle for the elderly is 10 minutes and that for adults is 15 minutes.

Fitted function for travel duration

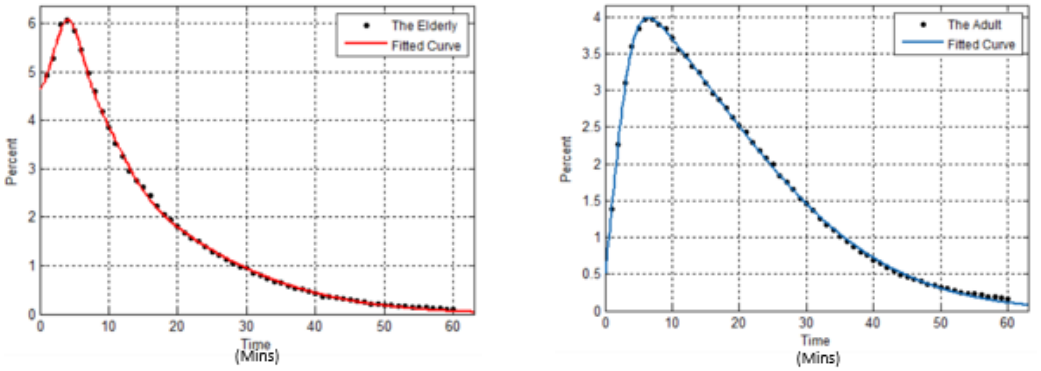
After testing all possible distribution functions, as shown in Figure 5.10, the Gaussian distribution function was chosen to find the travel duration distribution for the elderly and the adults. This function is best fitted to model travel duration.

Figure 5.10(a) is the fitted curve for the elderly's travel duration, and Figure 5.10(b) is the fitted curve for the adults' travel duration. The black dots represent the travel duration data, and the red and blue lines represent the respective fitted curves for the elderly and the adults, respectively.

In Figure 5.10(a), the Gaussian function (red line) represents the travel duration data of the elderly. Figure 5.10(b) also shows that the Gaussian function (in blue) represents the travel duration behavior of the adults.

Table 5.6 shows the parameters and goodness of fit. The R square values for the fitted Gaussian function and the fitted Weibull function are 0.9995 and 0.9996, respectively; and the SSE and RMSE values are very small. This proves that both fitted functions well fit the distribution of travel duration for the elderly and the adults, respectively.

From this, it can be concluded that (a) the travel duration for the elderly follows a Gaussian distribution, and (b) the travel duration for the adults follows a Gaussian distribution.



(a) The Elderly (b) The Adults

Figure 5. 10 The fitted curves of travel duration distribution

Table 5. 6 Fitted model of travel duration and goodness of fit

	Elderly	Adults
Model	$f(x)$ $= 1.447e^{(-(x-4.018/2.767)^2)}$ $+ 1.246e^{(-(x-6.835/6.394)^2)}$ $+ 5.199e^{(-(x+18.42/37.07)^2)}$	$f(x)$ $= -5.13e^{(-(x+1.819/4.222)^2)}$ $+ 5.024e^{(-(x+9.727/35.67)^2)}$
Goodness of fit		
SSE	0.09121	0.03649
R square	0.9995	0.9996
RMSE	0.04229	0.02599

5.3.4 Travel frequency

Definition of travel frequency

In this chapter, travel frequency is defined as the number of times a bus is taken within a given period, such as a whole day, morning, or afternoon. The travel frequency of the elderly was further compared with that in adults.

Based on the statistical analysis of the trip data, the travel frequency does not normally exceed 6 trips. This number was thus used as the criteria for the largest number of trips per day. Trip records

with more than 6 daily trips were removed as outliners accordingly.

Using the trip data of 243,711 elderly's smart card holders and the trip data of 1,865,378 adults smart card holders, a statistical description of the travel frequency characteristics of the two groups was obtained and illustrated in Figure 5.11. Figure 5.11(a) shows the statistical description of the travel frequency of the elderly, whereas Figure 5.11(b) shows the travel frequency of the adults. The travel frequency characteristics for both the elderly and adults are summarized in Table 5.7. The first row of Table 5.7(a) shows that the travel frequency of the elderly ranges from 1 to 6 times per day. The respective mean value and median of travel frequency are 1.81 and 2 times per day with the standard deviation of 1.02. Between the mean and median, we recommend the mean rather than the median because the outliners have been removed and the mean is more precise, to two decimal places. The second row of Table 5.7(a) shows that the travel frequency of the adults also ranged from 1 to 6 times per day. The respective mean value and median of travel frequency are 1.53 and 1 times per day, with a standard deviation of 0.78. From the above statistics, we can see that the elderly travel more times per day than the adults (1.81 vs 1.53), possibly because the adults take the subway to work and to social activities and thus travel by bus less frequently than the elderly .

As a further step, we compare the travel frequency of the two groups in terms of morning and afternoon (including evening) time periods. Table 5.7(b) shows that in the morning, the mean travel frequency of the elderly and the adults are 1.50 and 1.19, respectively; and those for the afternoon are 1.50 and 1.26, respectively, based on Table 5.7(c).

From these results, it is clearly that in the morning or afternoon the elderly's travel frequency by bus is higher than that of the adults (1.50 vs 1.19, respectively). One possible explanation is that the elderly may take a trip from home in the morning and also a return trip at that time. Thus, two morning trips take place, whereas an adult may travel to work in the morning and return home in the afternoon or evening, hence one morning trip and one afternoon trip take place. This situation is paralleled in the afternoon, but the time difference between the two events is smaller: the travel frequency of the elderly is higher than that of adults (1.50 vs 1.26, respectively). This difference can

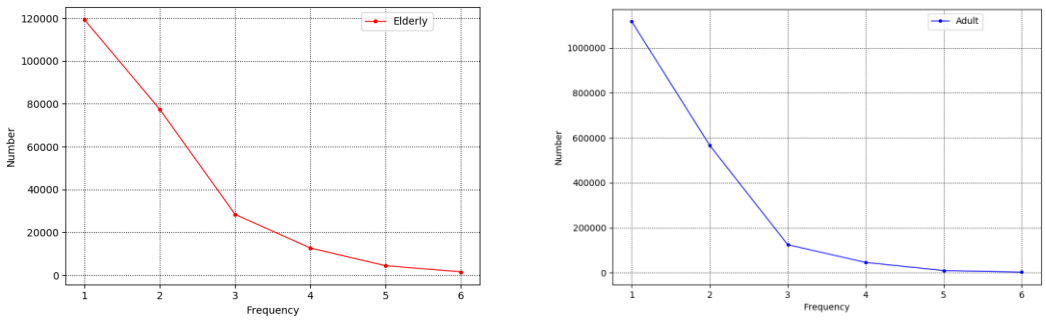
be explained by the different daily responsibilities of the two groups.

The travel frequency for the elderly is the same in the morning and afternoon (1.50 vs 1.50); however, the travel frequency of the adults is quite different in the morning and afternoon (1.19 vs 1.26, respectively). This may be because the adults may have more social activities after work in the evening, and thus travel more frequently in the afternoon and evening than in the morning.

Figure 5.11 shows that there is a different travel frequency trend for the elderly and the adults. Among the more than 243,000 the elderly' travel, only 119,000 travel once a day; more than 77,000 the elderly's travel twice a day. A high number of the elderly decrease their travel times from 1 to 3 times per day; and the number of the elderly who travel 3 times per day or more is decreasing.

Figure 5.11(b) shows the adults' trend of travel frequency. Among more than 1,865,000 adults' travel, more than 1,600,000 travel one or two times per day; 124,000 adults travel three times a day; and 45,000 adults travel four times a day. There is sharp decrease in the number of adults who travel 3 to 5 times per day; and the number of adults who travel 6 times a day or more is even lower.

The travel frequency of the elderly is as follows: (a) most the elderly travel once per day, on average each elderly person travels 1.81 times a day, which is more often than adults (mean, 1.53). (b) The travel frequency for the elderly population is the same in the morning and in the afternoon (1.50 vs 1.50). However, the behavior is quite different for adults (1.19 vs 1.26, respectively).



(a) The Elderly (b) The Adults

Figure 5. 11 Travel frequency analysis based on smart card data

Table 5. 7 Travel frequency of the elderly

(a) For whole day

	ID Card Number	Minimum	Maximum	Median	Mean	STD
Elderly	243,711	1	6	2	1.81	1.02
Adults	1,865,378	1	6	1	1.53	0.78

(b) For morning

	ID Card Number	Minimum	Maximum	Median	Mean	STD
Elderly	164,414	1	6	1	1.50	0.75
Adults	1,093,170	1	6	1	1.19	0.46

(c) For afternoon

	ID Card Number	Minimum	Maximum	Median	Mean	STD
Elderly	128,957	1	6	1	1.50	0.77
Adults	1,249,357	1	6	1	1.26	0.55

Fitted function for travel frequency

After testing all possible distribution functions, as shown in Figure 5.12, the Gaussian distribution function was chosen to find the travel frequency distribution for the elderly and the adults. This function is best fitted to model travel frequency.

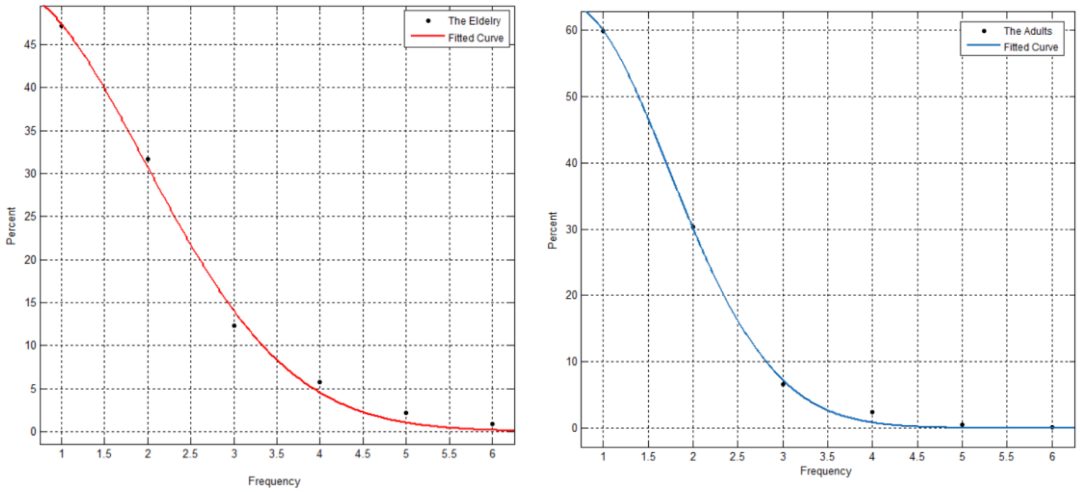
Figure 5.12(a) is the fitted curve for the elderly's travel frequency, and Figure 5.12(b) is the fitted curve for the adults' travel frequency. The black dots represent the travel frequency data, and the

red and blue lines represent the respective fitted curves for the elderly and the adults, respectively.

In Figure 5.12(a), the Gaussian function (red line) represents the travel frequency data of the elderly. Figure 5.12(b) also shows that the Gaussian function (in blue) represents the travel frequency behavior of the adults.

Table 5.8 shows the parameters and goodness of fit. The R square values for the fitted Gaussian function and the fitted Weibull function are 0.9959 and 0.9989, respectively; and the SSE and RMSE values are not very large. This proves that both fitted functions well fit the distribution of travel duration for the elderly and the adults, respectively.

From this, it can be concluded that (a) the travel frequency for the elderly follows a Gaussian distribution, and (b) the travel frequency for the adults also follows a Gaussian distribution.



(a) The Elderly (b) The Adults

Figure 5. 12 The fitted curves of travel frequency distribution

Table 5. 8 Fitted model of travel frequency and goodness of fit

	Elderly	Adults
Model	$f(x) = 52.06e^{(-(x-0.2703/2.387)^2)}$	$f(x) = 63.99e^{(-(x-0.5783/1.641)^2)}$
	Goodness of fit	
SSE	7.225	3.231
R square	0.9959	0.9989
RMSE	1.552	1.038

5.4 Findings and discussions

Travel distance

The elderly like to travel by bus; most travel approximately 1 km, and the median is 4 km, whereas the median for adults is 5 km. The number of the elderly who travel more than 12 km is very small. Furthermore, the travel distance distribution of the elderly follows an exponential function, which is different from the travel distance distribution of the adults, which follows a Gaussian function.

The above findings can be used directly for policymaking and management related to the elderly. For example, with the findings that most elderly travel approximately 1 km with the median of 4 km by bus, Beijing city is recommended to initiate a short-distance bus service policy for the elderly (such as a community bus service), and the bus service distance for most of this kind of special service can be set as 1 km. Furthermore, the finding that the travel distance distribution of the elderly follows an exponential function tells us the distribution of the travel distance. This finding provides the scientific basis for the short-distance bus service scheduling for the elderly, that is, the percentage of bus service for a particular distance should follow an exponential distribution, with 1 km as the maximum of this distribution.

With the characteristics of smart card data and a data-driven approach, it is possible to provide such travel distance information in real time to the bus service provider to optimize the service. As a result, the distance of services for the elderly could be -adjusted dynamically based on the distance

statistics 1 hour or 1 day earlier. Such precise short-distance bus service for the elderly based on smart card analytics will not only provide on-demand service to the elderly, but will also maximize profit for companies that provide such service.

Travel duration

Most elderly use buses for 4 minutes and most adults for 6 minutes. The median travel duration on public transport vehicles by the elderly is 10 minutes and that for adults is 15 minutes. Adults travel duration is larger than the travel duration of the elderly. Travel duration for the elderly and the adults follow a Gaussian distribution.

Similar to the travel distance results, the findings of travel duration for the elderly can also be used for policymaking and management related to the elderly. For example, with the findings that most elderly travel for 4 minutes daily with the median of 10 minutes by bus in Beijing, the municipal government is recommended to consider providing a special bus service for the elderly, with the considerations that most of the elderly will take a bus service for 4 minutes with the median of 10 minutes. More precisely, the travel time distribution of the special bus service can be set according to the discovered regularity in the Gaussian distribution of the elderly's travel duration. The requirement of the elderly on their travel duration can thus be best satisfied statistically, and the company can make the best benefit accordingly.

Travel departure and arrival time

Regarding the travel departure and arrival times, there is a peak in the morning at 9:00 AM and no clear peak in the afternoon, for the elderly's departure time; whereas for adults, there is a peak in the morning at 8:00 AM and three small departure peaks between 5:13 and 6:18 PM. Although the morning peak of departure trips for the elderly (9:00 AM) is 1 hour later than that for the adults, mainly workers (8:00 AM), there is a strong overlap for 2 hours between the distribution of the two peak periods. This is no peak overlap for the afternoon departure time between the elderly and the adults.

The above findings reflect the fact that the trips of the elderly also contribute to the Beijing's overall morning traffic congestion. It is therefore recommended that Beijing may try to reduce this traffic congestion problem by initiating a new bus price policy for the elderly. For example, with the new policy, the elderly can enjoy a greater bus fare discount if they travel by bus beyond the morning adults' peak hour period (7:00 to 9:00 AM). This will not only alleviate traffic congestion for the workers in the morning, but the elderly will also be able to enjoy a safer and more comfortable travel environment with fewer bus users from the adults cohort.

Furthermore, spatial distribution analysis indicates a strong spatial location overlap of the bus stops used by both the elderly and the adults during the morning peak of the adults (8:00 AM). This is especially true for the area between the 2nd and 4th ring roads, and the situation is even worse for the northwest area of Beijing. Regarding this traffic congestion problem, a location-dependent bus fare policy can be implemented. For example, with the new policy, there will be no bus fare discount if an elderly person travels by bus inside the area between the 2nd and 4th ring roads of Beijing during the peak period 7:00 to 9:00 AM, and a greater fare discount can be enjoyed outside this area during the peak period.

Travel frequency

Regarding travel frequency, in general, most the elderly travel once per day, and on average, each person travels 1.81 times a day, which is more than that of adults, whose average travel time is 1.53 times a day. The travel frequency for the elderly is the same in both morning and afternoon (1.50). Travel frequency for the elderly and the adults follow a Gaussian distribution.

The findings on elderly's travel frequency is very important for transport scheduling and public service related to the elderly. For example, the finding that most the elderly travel once per day and this is evenly distributed for morning and afternoon statistically can guide us for the corresponding transport scheduling for the elderly. In the future autonomous driving bus service for the elderly, the same number of buses should be scheduled for morning and afternoon for the elderly, with the consideration that each elderly person normally travels one time per day. A similar scheduling

arrangement could also be applied to public facilities oriented to the elderly, including medical services, shops, and recreation facilities.

5.5 Conclusion

In response to the worldwide challenge of aging populations, this chapter was conducted to investigate the mobility behavior of the elderly in Beijing, including departure and arrival time, travel distance, duration, and frequency. In the era of big data, a data-driven approach was used for this chapter. Smart card data were used for the analysis, which can provide travel behavior analysis with data in near–real time, a large sampling size (millions), precise latitude and longitude coordinates, and high data updating frequency in hourly rate. Two methods were used for this chapter: (a) quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent presentation of such behavior graphically, and (b) the discovery of the distribution function of the travel characteristics.

This chapter makes a series of important findings on the elderly's mobility behaviors in a megacity: (a) *travel distance*: most travel approximately 1 km; the median is 4 km, which is shorter than in adults; it follows an exponential function; (b) *travel duration*: most travel for 4 minutes, which is half that of adults; it follows a Gaussian distribution; (c) *travel departure time*: a peak is seen at 9:00 AM (compared with 8:00 AM for adults), and no clear peak is seen in the afternoon; and (d) *travel frequency*: most the elderly travel once per day by bus, which is same with adults. These findings can make a significant contribution to smart city transport planning, management, and services in light of the world's aging population.

Chapter 6 Conclusions

This thesis project of three and a half years was devoted to analyzing the spatiotemporal distribution and mobility behavior of the elderly using spatial big data. Three achievements were realized: a) a series of new data analytics methods; b) findings regarding the spatiotemporal distribution and mobility behavior of the elderly; and c) identification of potential applications of the findings.

6.1 A series of new data analytics methods

Voronoi construction based on an integrated spatial clustering method

Voronoi construction method based on integrated spatial clustering method was developed for city partition. The integrated clustering method can detect clusters in datasets with multiple densities and shapes. Two improvements were made to classic clustering methods: a) cluster number could be estimated automatically, and b) one parameter was required. With these improvements, multiple densities and shapes of clusters could be detected effectively. The clustering method was also scalable for different kinds of dataset. Voronoi diagram is constructed according to the center points of each cluster. It is used to analyze the spatial distribution of the elderly.

A data-driven framework for analyzing the spatial distribution of the elderly

The framework of the method was designed to systematically answer the following three questions: a) where do the elderly live? b) why the elderly distribution like this? c) where do they go frequently? The framework includes the following key methods: (a) the integrated clustering method is used to identify the home locations of the elderly based on bus stop locations, (b) a Voronoi diagram-based method of partitioning a city into regions that reflect the natural clustering of the elderly's living spaces, and (c) an improved clustering method to cluster bus stops based on the elderly's flow at stops and to create a Voronoi diagram based on the cluster centers.

A PoI-based elderly livability index model

A PoI-based elderly livability index was developed to measure and explain the factors that are important for elderly citizens when choosing their home location, and to rationalize the spatial distribution of the elderly population. The index emphasizes public facilities, with restaurants, parks, hospitals, shops and bus stops identified as the key factors in the model. The model generates an indicator of an area's PoI-based elderly livability for the elderly.

A spatial connectivity approach to travel destination analysis

Spatial connectivity was used to analyze the elderly's travel behavior between any two locations, such as from home to hospital. Connectivity indicates the links from elderly citizens' home areas to the areas where they undertake activities.

A quasi-gravity model of connectivity between regions

A quasi-gravity model was developed and confirmed to be valid for the relationship between the spatial connectivity between any two regions in terms of network strength and the PoI-based elderly livability index of the regions.

A framework for analyzing the mobility behavior of the elderly

A framework composed of data capture, data preprocessing (including data cleaning), and mobility analytics was developed based on a) the traveling distance of the elderly, b) their departure and arrival time, c) the travel frequency of the elderly, and d) the travel duration of the elderly. Two data-driven methods were used: a) quantitative analysis of spatiotemporal travel behavior by estimating the parameters of travel patterns and the subsequent presentation of such behavior graphically, and b) discovery of the distribution function of the travel characteristics.

6.2 Important scientific findings

A series of findings regarding the spatial distribution and mobility behavior of the elderly were revealed by applying the proposed method to smart card data from Beijing.

Spatial distribution of the elderly

Three findings were obtained regarding the spatial distribution of the elderly: a) the spatial distribution of the elderly shows clear clustering characteristics; b) the spatial distribution of the elderly has a strong relationship with the provision of public service facilities, such as restaurants and hospitals; and c) the connectivity of each pair of regions is related to the distribution of public facilities in the connected regions. The spatial connectivity between any two regions in terms of network strength and the PoI-based elderly livability index of the different regions follows a quasi-gravity model.

Mobility behavior of the elderly

Four findings regarding elderly mobility behaviors in a megacity were obtained. (a) Travel distance: most travels taken by the elderly are approximately 1 km and the median distance is 4 km, which is shorter than the distance traveled by the adults. The travel distance of the elderly follows an exponential function, which is different from the travel distance distribution of the adults, which follows a Gaussian function. (b) Travel duration: most of the elderly travel for 4 minutes, which is half the time of the adults' travel; travel duration follows a Gaussian distribution. (c) Travel departure time: there is a morning peak at 9:00 am (compared with 8:00 am for adults) and no clear peak in the afternoon. (d) Travel frequency: most elderly travel once per day, which is less frequently than adults.

6.3 Significance and potential applications of the findings

The significance of this thesis lies in the series of methods developed and its important findings regarding the spatial distribution patterns and mobility behavior of the elderly in megacities. These findings add new knowledge to the field and the new methods can be widely applied for urban transport planning, management and services for the ageing population.

The findings for elderly mobility can be applied directly to transport policymaking and management. For example, given the finding that most elderly travel approximately 1 km with a median distance of 4 km by bus, it is recommended that cities initiate a short-distance bus service policy for the elderly, and the bus service distance for most of these special services can be set as 1 km.

The finding that the travel distance distribution of the elderly follows an exponential function provides a scientific basis for the scheduling of short-distance bus services for the elderly. The percentage of bus services for a particular distance should follow an exponential distribution, with 1 km as the maximum of this distribution.

Smart card data make it possible to provide travel distance information in real time to the bus service provider to optimize its service. As a result, the distance of special services for the elderly could be finely adjusted dynamically based on the distance statistics one hour earlier. A precisely timetabled short-distance bus service for the elderly based on smart card analytics could not only provide an on-demand service for the elderly, but also maximize profit for the companies providing such a service.

The findings on the travel departure time of the elderly and the adults show that the trips made by the elderly contribute to the overall morning traffic congestion in the city. It is therefore recommended that the city tries to reduce the traffic congestion problem by initiating a new bus fare policy for the elderly. For example, the elderly could enjoy a greater bus fare discount if they travel outside the morning adult peak period (7:00 am to 9:00 am). Not only will this alleviate traffic congestion for workers in the morning, but the elderly will also be able to enjoy a safer and more comfortable travel environment with fewer passengers from the adult cohort.

Furthermore, the spatial distribution analysis indicates that there is a strong spatial location overlap of the bus stops used by the elderly and the adults during the morning peak of adult journeys (8:00 am). To address the traffic congestion problem, a location-dependent bus fare policy could be implemented. For example, there could be no fare discount if an elderly person travels by bus inside

the overlapping area during the peak period, and a greater fare discount outside the overlapping area.

The findings regarding the travel frequency of the elderly can be used for transport scheduling and public services related to the elderly. For example, the finding that most of the elderly travel once a day and this travel is evenly distributed in the morning and afternoon can guide transport scheduling for the elderly. In the future, when scheduling autonomous bus services for the elderly, the same number of buses should be scheduled for the morning and afternoon, with the consideration that each elderly person normally travels once daily. A similar scheduling arrangement could also be applied to elderly-oriented public facilities, including medical services, shops and recreation facilities.

6.4 Future work

This thesis investigated the spatiotemporal and mobility behavior characteristics of the elderly using smart card data. The following two issues that were not investigated due to time and data limitations could be studied in the future.

First, smart card data could be integrated with survey data such as a questionnaire for the elderly. Smart card data have unique advantages, such as large volume and detailed travel information. However, smart card data do not capture information about the card holder such as personal details and travel purpose. For this reason, questionnaire data should be considered together with smart card data. This would provide details such as age, gender and travel purpose for the analysis.

Second, because of the limitations of smart card data, this thesis only considered the bus transport mode when studying the spatial distribution of the elderly. Other transport modes, such as walking, bicycles, mass transit railways and private cars, could be studied when relevant data sets are available.

References

Adin, A., Lee, D., Goicoa, T., & Ugarte, M. D. (2018). A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. *Statistical methods in medical research*, 0962280218767975.

Ahern, A., & Hine, J. (2012). Rural transport–Valuing the mobility of older people. *Research in transportation economics, 34*(1), 27-34.

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention, 41*(3), 359-364.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). *OPTICS: ordering points to identify the clustering structure.* Paper presented at the ACM Sigmod Record.

Atluri, G., Karpatne, A., & Kumar, V. (2017). Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Computing Surveys, 1*(1). doi:10.1145/3161602

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., . . . Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports, 734*, 1-74.

Barry, J. J., Newhouser, R., Rahbee, A., & Sayeda, S. (2002). Origin and destination estimation in New York City with automated fare system data. *Transportation Research Record, 1817*(1), 183-187.

Beijing Committee on Aging. (2018). *White papers of development of aging service and care system construction in Beijing*

Birant, D., & Kut, A. (2006). An algorithm to discover spatial–temporal distributions of physical seawater characteristics and a case study in Turkish seas. *Journal of marine science and technology, 11*(3), 183-192.

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering, 60*(1), 208-221.

Böcker, L., & Thorsson, S. (2014). Integrated weather effects on cycling shares, frequencies, and durations in Rotterdam, the Netherlands. *Weather, climate, and society, 6*(4), 468-481.

Böcker, L., van Amen, P., & Helbich, M. (2017). Elderly travel frequencies and transport mode choices in Greater Rotterdam, the Netherlands. *Transportation, 44*(4), 831-852.

Boschmann, E. E., & Brady, S. A. (2013). Travel behaviors, sustainable mobility, and transit-oriented developments: a travel counts analysis of older adults in the Denver, Colorado metropolitan area. *Journal of Transport Geography, 33*, 1-11.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing, 93*(1), 79-95.

Brunsdon, C., Corcoran, J., & Higgs, G. (2007). Visualising space and time in crime patterns: A comparison of methods. *Computers, Environment and Urban Systems, 31*(1), 52-75.

Burtless, G. (2013). The impact of population aging and delayed retirement on workforce productivity. *SSRN*.

Byers, S., & Raftery, A. E. (1998). Nearest-neighbor clutter removal for estimating features in spatial point processes. *Journal of the American Statistical Association, 93*(442), 577-584.

Chen, Y., Liu, J., & Ge, H. (1999). Pattern characteristics of foreshock sequences. In *Seismicity Patterns, their Statistical Significance and Physical Meaning* (pp. 395-408): Springer.

Cheng, L., Caset, F., De Vos, J., Derudder, B., & Witlox, F. (2019). Investigating walking accessibility

to recreational amenities for elderly people in Nanjing, China. *Transportation Research Part D: Transport and Environment, 76*, 85-99.

Cheng, L., Chen, X., Yang, S., Cao, Z., De Vos, J., & Witlox, F. (2019). Active travel for active ageing in China: The role of built environment. *Journal of Transport Geography, 76*, 142-152.

Cheng, T., Haworth, J., Anbaroglu, B., Tanaksaranond, G., & Wang, J. (2014). Spatiotemporal data mining. In *Handbook of Regional Science* (pp. 1173-1193): Springer.

Cheng, T., & Li, Z. (2004). *A hybrid approach to detect spatial-temporal outliers.* Paper presented at the Proceedings of the 12th International Conference on Geoinformatics Geospatial Information Research.

Cheng, T., & Li, Z. (2006). A multiscale approach for spatio-temporal outlier detection. *Transactions in GIS, 10*(2), 253-263.

Cheng, T., & Wicks, T. (2014). Event detection using Twitter: a spatio-temporal approach. *PloS one, 9*(6), e97807.

Choo, S., Sohn, D., & Park, M. (2016). Mobility characteristics of the elderly: A case for Seoul Metropolitan Area. *KSCE Journal of Civil Engineering, 20*(3), 1023-1031.

Collia, D. V., Sharp, J., & Giesbrecht, L. (2003). The 2001 national household travel survey: A look into the travel patterns of older Americans. *Journal of safety research, 34*(4), 461-470.

Cressie, N., & Wikle, C. K. (2015). *Statistics for Spatio-Temporal Data*: John Wiley & Sons.

Cui, J., Loo, B. P., & Lin, D. (2017). Travel behaviour and mobility needs of older adults in an ageing and car-dependent society. *International Journal of Urban Sciences, 21*(2), 109-128.

Delmelle, E., Casas, I., Rojas, J. H., & Varela, A. (2013). Spatio-temporal patterns of dengue fever in Cali, Colombia. *International Journal of Applied Geospatial Research (IJAGR), 4*(4), 58-75.

Delmelle, E., Dony, C., Casas, I., Jia, M., & Tang, W. (2014). Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science, 28*(5), 1107-1127.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological), 39*(1), 1-22.

Di Martino, F., & Sessa, S. (2011). The extended fuzzy C-means algorithm for hotspots in spatio-temporal GIS. *Expert Systems with Applications, 38*(9), 11829-11836.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 349-362.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*: Chapman and Hall/CRC.

Erwig, M. (2000). The graph Voronoi diagram with applications. *Networks, 36*(3), 156-163.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* Paper presented at the Proceedings of the 2nd International Conference on Knowledge and Discovery and Data Mining.

Feng, J. (2017). The influence of built environment on travel behavior of the elderly in urban China. *Transportation Research Part D: Transport and Environment, 52*, 619-633.

Findlay, A. M., Stockdale, A., Findlay, A., & Short, D. (2001). Mobility as a driver of change in rural Britain: an analysis of the links between migration, commuting and travel to shop patterns. *International Journal of Population Geography, 7*(1), 1-15.

Fortune, S. (1987). A sweepline algorithm for Voronoi diagrams. *Algorithmica, 2*(1-4), 153.

Fränti, P., & Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence, 48*(12), 4743-4759.

Glaz, J., Naus, J. I., Wallenstein, S., Wallenstein, S., & Naus, J. I. (2001). *Scan Statistics*: Springer.

Goins, R. T., Williams, K. A., Carter, M. W., Spencer, S. M., & Solovieva, T. (2005). Perceived barriers to health care access among rural older adults: a qualitative study. *The Journal of Rural Health, 21*(3), 206-213.

Gomide, J., Veloso, A., Meira Jr, W., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011). *Dengue surveillance based on a computational model of spatio-temporal locality of Twitter.* Paper presented at the Proceedings of the 3rd International Web Science Conference.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *nature, 453*(7196), 779.

Guha, S., Rastogi, R., & Shim, K. (1998). *CURE: an efficient clustering algorithm for large databases.* Paper presented at the ACM Sigmod Record.

Guo, S., Song, C., Pei, T., Liu, Y., Ma, T., Du, Y., . . . Peng, Y. (2019). Accessibility to urban parks for elderly residents: Perspectives from mobile phone data. *Landscape and Urban Planning, 191*, 103642.

Hahn, J.-S., Kim, H.-C., Kim, J.-K., & Ulfarsson, G. F. (2016). Trip making of older adults in Seoul: Differences in effects of personal and household characteristics by age group and trip purpose. *Journal of Transport Geography, 57*, 55-62.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*: Elsevier.

He, S. Y., Cheung, Y. H., & Tao, S. (2018). Travel mobility and social participation among older people in a transit metropolis: A socio-spatial-temporal perspective. *Transportation research part A: policy and practice, 118*, 608-626.

Hinneburg, A., & Gabriel, H.-H. (2007). *Denclue 2.0: Fast clustering based on kernel density estimation.* Paper presented at the International Symposium on Intelligent Data Analysis.

Hu, X., Wang, J., & Wang, L. (2013). Understanding the travel behavior of elderly people in the developing country: a case study of Changchun, China. *Procedia of Social and Behavioral Sciences, 96*, 873-880.

Hu, Y., Wang, F., Guin, C., & Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography, 99*, 89-97.

Jacquez, G. M. (1996). A k nearest neighbour test for space–time interaction. *Statistics in medicine, 15*(18), 1935-1949.

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data, 3*(2), 208-219.

Jiang, S., Ferreira Jr, J., & Gonzalez, M. C. (2012). *Discovering urban spatial-temporal structure from human activity patterns.* Paper presented at the Proceedings of the ACM SIGKDD International Workshop on Urban Computing.

Johnston, K., Ver Hoef, J. M., Krivoruchko, K., & Lucas, N. (2001). *Using ArcGIS Geostatistical Analyst* (Vol. 380): Esri Redlands.

Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *computer, 32*(8), 68-75.

Kim, S. (2011). Assessing mobility in an aging society: Personal and built environment factors associated with older people's subjective transportation deficiency in the US. *Transportation research part F: traffic psychology and behaviour, 14*(5), 422-429.

Kisilevich, S., Mansmann, F., Nanni, M., & Rinzivillo, S. (2009). *Data Mining and Knowledge Discovery Handbook*: Springer.

Knox, E., & Bartlett, M. (1964). The detection of space-time interactions. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 13*(1), 25-30.

Knox, G. (1963). Detection of low intensity epidemicity: application to cleft lip and palate. *British journal of preventive & social medicine, 17*(3), 121.

Krisp, J. M., Peters, S., & Burkert, F. (2013). Visualizing crowd movement patterns using a directed kernel density estimation. In *Earth Observation of Global Changes (EOGC)* (pp. 255-268): Springer.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods, 26*(6), 1481-1496.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 164*(1), 61-72.

Kulldorff, M. (2018). SaTScan v9.6: Software for the spatial, temporal, and space-time scan statistics. *Information Management Services Inc.*

Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R., & Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS medicine, 2*(3), e59.

Kulldorff, M., & Hjalmars, U. (1999). The Knox method and other tests for space-time interaction. *Biometrics, 55*(2), 544-552.

Lee, J., Gong, J., & Li, S. (2017). Exploring spatiotemporal clusters based on extended kernel estimation methods. *International Journal of Geographical Information Science, 31*(6), 1154-1177.

Lin, T. G., Xia, J. C., Robinson, T. P., Goulias, K. G., Church, R. L., Olaru, D., . . . Han, R. (2014). Spatial analysis of access to and accessibility surrounding train stations: A case study of accessibility for the elderly in Perth, Western Australia. *Journal of Transport Geography, 39*, 111-120.

Liu, W., Lu, H., Sun, Z., & Liu, J. (2017). Elderly's travel patterns and trends: The empirical analysis of Beijing. *Sustainability, 9*(6), 981.

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems, 53*, 19-35.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research, 27*(2 Part 1), 209-220.

McDonald, N. C. (2008). Children's mode choice for the school trip: the role of distance and school location in walking to school. *Transportation, 35*(1), 23-35.

Miller, H. J., & Han, J. (2009). *Geographic Data Mining and Knowledge Discovery*: CRC Press.

Mohamed, K., Côme, E., Oukhellou, L., & Verleysen, M. (2016). Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems, 18*(3), 712-728.

Moniruzzaman, M., Chudyk, A., Paez, A., Winters, M., Sims-Gould, J., & McKay, H. (2015). Travel behavior of low income older adults and implementation of an accessibility calculator. *Journal of transport & health, 2*(2), 257-268.

Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space - time Kernel Density Estimation and Scan Statistics. *Transactions in GIS, 14*(3), 223-239.

Napier, G., Lee, D., Robertson, C., & Lawson, A. (2018). A Bayesian space–time model for clustering areal units based on their disease trends. *Biostatistics*.

Neill, D. B. (2006). Detection of spatial and spatio-temporal clusters. In *Tech Rep CMU-CS-06-142, PhD*

thesis: Carnegie Mellon University.

Páez, A., Scott, D., Potoglou, D., Kanaroglou, P., & Newbold, K. B. (2007). Elderly mobility: demographic and spatial analysis of trip making in the Hamilton CMA, Canada. *Urban Studies, 44*(1), 123-146.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications, 36*(2), 3336-3341.

Pei, T., Jasra, A., Hand, D. J., Zhu, A.-X., & Zhou, C. (2009). DECODE: a new method for discovering clusters of different densities in spatial data. *Data Mining and Knowledge Discovery, 18*(3), 337.

Pei, T., Zhou, C., Zhu, A.-X., Li, B., & Qin, C. (2010). Windowed nearest neighbour method for mining spatio-temporal clusters in the presence of noise. *International Journal of Geographical Information Science, 24*(6), 925-948.

Plazinić, B. R., & Jović, J. (2018). Mobility and transport potential of elderly in differently accessible rural areas. *Journal of Transport Geography, 68*, 169-180.

Raykov, Y. P., Boukouvalas, A., Baig, F., & Little, M. A. (2016). What to do when K-means clustering fails: a simple yet principled alternative algorithm. *PloS one, 11*(9).

Ripepe, M., Piccinini, D., & Chiaraluce, L. (2000). Foreshock sequence of September 26th, 1997 Umbria-Marche earthquakes. *Journal of Seismology, 4*(4), 387-399.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492-1496.

Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data Mining and Knowledge Discovery Handbook* (pp. 321-352): Springer.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., . . . Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing, 267*, 664-681.

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*: John Wiley & Sons.

Shao, F., Sui, Y., Yu, X., & Sun, R. (2019). Spatio-temporal travel patterns of elderly people–A comparative study based on buses usage in Qingdao, China. *Journal of Transport Geography, 76*, 178-190.

Shekhar, S., Evans, M. R., Kang, J. M., & Mohan, P. (2011). Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*(3), 193-214.

Shekhar, S., Jiang, Z., Ali, R. Y., Eftelioglu, E., Tang, X., Gunturi, V., & Zhou, X. (2015). Spatiotemporal data mining: a computational perspective. *ISPRS International Journal of Geo-Information, 4*(4), 2306-2338.

Shekhar, S., Vatsavai, R. R., & Celik, M. (2008). Spatial and spatiotemporal data mining: Recent advances. In *Data Mining: Next Generation Challenges and Future Directions* (pp. 1-34): Chapman and Hall/CRC.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis* (Vol. 26): CRC Press.

Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics, 6*(10), 818.

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science, 327*(5968), 1018-1021.

Sun, L., Lee, D.-H., Erath, A., & Huang, X. (2012). *Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system.* Paper presented at the Proceedings of

the ACM SIGKDD international workshop on urban computing.

Szeto, W., Yang, L., Wong, R., Li, Y., & Wong, S. (2017). Spatio-temporal travel characteristics of the elderly in an ageing society. *Travel behaviour and society, 9*, 10-20.

Takahashi, K., Kulldorff, M., Tango, T., & Yih, K. (2008). A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics, 7*(1), 14.

Takahashi, K., Yokoyama, T., & Tango, T. (2013). FleXScan v3. 1. 2: Software for the Flexible Scan Statistic. *National Institute of Public Health, Japan*.

Tang, J., Chang, Y., & Liu, H. (2014). Mining social media with social theories: a survey. *Acm Sigkdd Explorations Newsletter, 15*(2), 20-29.

Tango, T., & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics, 4*(1), 11.

Tango, T., Takahashi, K., & Kohriyama, K. (2011). A space–time scan statistic for detecting emerging outbreaks. *Biometrics, 67*(1), 106-115.

Titheridge, H., Achuthan, K., Mackett, R., & Solomon, J. (2009). Assessing the extent of transport social exclusion among the elderly. *Journal of Transport and Land Use, 2*, 31-48.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234-240.

Truong, L. T., & Somenahalli, S. V. (2015). Exploring frequency of public transport use among older adults: A study in Adelaide, Australia. *Travel behaviour and society, 2*(3), 148-155.

Uncu, O., Gruver, W. A., Kotak, D. B., Sabaz, D., Alibhai, Z., & Ng, C. (2006). *GRIDBSCAN: GRId density-based spatial clustering of applications with noise.* Paper presented at the 2006 IEEE International Conference on Systems, Man and Cybernetics.

United Nations. (2019a). *World Population Prospects 2019: Highlights (ST/ESA/SER.A/423)*.

United Nations. (2019b). *World Urbanization Prospects 2018: Highlights (ST/ESA/SER.A/421)*.

United Nations. (2020). *World Population Ageing 2019 (ST/ESA/SER.A/444)*.

Van den Berg, P., Arentze, T., & Timmermans, H. (2011). Estimating social travel demand of senior citizens in the Netherlands. *Journal of Transport Geography, 19*(2), 323-331.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). *Spatiotemporal data mining in the era of big spatial data: algorithms and applications.* Paper presented at the Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data.

Wang, J., Kong, X., Xia, F., & Sun, L. (2019). Urban Human Mobility: Data-Driven Modeling and Prediction. *Acm Sigkdd Explorations Newsletter, 21*(1), 1-19.

Wang, M., Wang, A., & Li, A. (2006). *Mining spatial-temporal clusters from geo-databases.* Paper presented at the International Conference on Advanced Data Mining and Applications.

Wang, W., Yang, J., & Muntz, R. (1997). *STING: A statistical information grid approach to spatial data mining.* Paper presented at the VLDB.

Wei, Q., She, J., Zhang, S., & Ma, J. (2018). Using individual GPS trajectories to explore foodscape exposure: A case study in Beijing metropolitan area. *International journal of environmental research and public health, 15*(3), 405.

Wong, R., Szeto, W., Yang, L., Li, Y., & Wong, S. (2018). Public transport policy measures for improving elderly mobility. *Transport policy, 63*, 73-79.

Wong, R., Szeto, W., Yang, L., Li, Y. C., & Wong, S. (2017). Elderly users' level of satisfaction with

public transport services in a high-density and transit-oriented city. *Journal of transport & health, 7*, 209-217.

Xia, X., & Guan, H. (2013). Travel Survey and Analyses of the Elderly in Beijing (in Chinese). *Urban Transport of China, 11*(5), 1-9.

Yang, L. (2018). Modeling the mobility choices of older people in a transit-oriented city: Policy insights. *Habitat international, 76*, 10-18.

Yuan, Y., Yang, M., Wu, J., Rasouli, S., & Lei, D. (2019). Assessing bus transit service from the perspective of elderly passengers in Harbin, China. *International Journal of Sustainable Transportation, 13*(10), 761-776.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). *BIRCH: an efficient data clustering method for very large databases.* Paper presented at the ACM Sigmod Record.

Zhang, Y., Yao, E., Zhang, R., & Xu, H. (2019). Analysis of elderly people's travel behaviours during the morning peak hours in the context of the free bus programme in Beijing, China. *Journal of Transport Geography, 76*, 191-199.

Zhao, P., Liu, X., Shi, W., Jia, T., Li, W., & Chen, M. (2018). An empirical study on the intra-urban goods movement patterns using logistics big data. *International Journal of Geographical Information Science*, 1-28.

Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(3), 29.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., & Schmitt, G. (2016). Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PloS one, 11*(2), e0149222.

Zhou, Z., & Matteson, D. S. (2015). *Predicting ambulance demand: A spatio-temporal kernel approach.* Paper presented at the Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Zou, Q., Yao, X., Zhao, P., Wei, H., & Ren, H. (2018). Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. *Transportation, 45*(3), 919-944.