Chapter 7

Public Transportation Big Data Mining and Analysis

Xiaolei Ma and Xi Chen
School of Transportation Science and Engineering, Beihang University, Beijing, People's Republic of China

Chapter Outline

7.1 INTRODUCTION

As city populations, number of motor vehicles, and residents' travel demands increase, urban road traffic flows are becoming ever more saturated with frequent urban traffic congestion. Urban transportation has become a problem that needs to be addressed urgently during the development process of many large- and medium-sized cities throughout the world. Based on practical experience accumulated worldwide with respect to urban transportation, prioritizing the

Data-Driven Solutions to Transportation Problems. https://doi.org/10.1016/B978-0-12-817026-7.00007-2

development of urban public transportation is an effective solution to the urban transportation problem. In recent years, the strategic position of public transportation in urban transportation has become increasingly critical, with an urgent need for constructing urban public transportation systems with high service levels and operating efficiencies to satisfy the development demand of urban transportation.

In recent years, with the emergence and continuous advancement of concepts such as intelligent transportation and smart cities, the urban transportation industry has established a relatively mature data acquisition and technology application foundation. With the advent of the big data era, massive, multisource spatial and temporal data, which are represented by cell phone data, floating car data, and social networks, are continuously employed in urban planning practice.

Manual survey data have been the primary data source for the analysis of the demand-and-supply relationship of public transportation systems and their operation and planning. Comprehensive spatial and temporal public transportation ridership distribution data are difficult to obtain by conventional transport survey methods. Conventional survey methods often collect residents' travel information via questionnaires or queries. Residents' travel data acquired using these methods generally do not satisfy the quantity and quality requirements for sample selection. These data have a series of problems, such as small survey sample sizes, low accuracy, high cost, low temporal effectiveness, long survey cycles, and difficulties in later-stage data processing. In addition, these data do not reflect the dynamic variation characteristics of passengers' travel demand.

Big data, urban public transportation smart cards, and satellite positioning technology, as well as informationalized survey and statistical methods for ridership, provide convenience for passengers. Massive data are generated during the public transportation operation process. These data are highly continuous, have wide coverage, contain comprehensive information, and update dynamically and quickly; thus, they have high application value to public transportation operation and planning.

During the urban public transportation operation process, the public transportation information infrastructure accumulates massive static and dynamic data in addition to providing transport information services. In this chapter, case studies are provided using data from the Beijing public transportation system to introduce the application of public transportation big data to practical problems. First, the basic types of public transportation data are introduced briefly. The following data types are commonly employed.

(1) Global positioning system (GPS) data

GPS data refer to the vehicle movement information acquired in real time by the GPS sensors installed on public transportation vehicles, such as taxis, buses, and subways.

(2) Ridership data

Urban residents choose different means of public transportation for their trips. The operational data of these means of public transportation constitute ridership data. Bus and subway ridership data are acquired by recording the usage of public transportation smart cards. These ridership data have high analytical value and can be used to monitor population movements, evaluate the urban public transportation system, and study the behavior of passengers.

(3) Mobile terminal data

Cell phones and tablet computers (e.g., iPad) are the mainstream mobile communication tools in today's society. Users can search for various types of transport information in real time with their cell phones and tablet computers. This mode of enquiry provides various types of data. Many public transportation systems (including bus routes) have launched social networking services to gather residents' feedback in real time to make timely adjustments and improvements.

(4) Video surveillance data

Video surveillance technology has been applied extensively in transport management. Massive video data collected by video surveillance equipment each day record urban residents' travel zones, which form a virtual "map" of residents' trips within a physical city in digital space. Cameras are also installed by the boarding and alighting doors of buses, and are primarily employed to capture video images of passengers. Data acquisition personnel determine the number of boarding and alighting passengers by analyzing and processing continuous video images using software.

(5) Wireless-fidelity (Wi-Fi) data

Existing dynamic urban public transportation monitoring systems rely on a third-generation (3G) network, which had shortcomings such as slow data transmission and high communication costs. A Wi-Fi-based on-board data transmission scheduling strategy for public transportation vehicles can reduce the total energy consumption of vehicles and data transmission time. Wi-Fi-based wireless transmission of on-board data of a public transportation vehicle refers to the transmission of updated data that are required by the vehicle via the File Transfer Protocol (FTP) by an on-board embedded device that connects to the server via the local Wi-Fi network.

This chapter is organized as follows: Section 7.2 introduces a quality control method for public transportation big data via case studies using data from the Beijing public transportation system. Section 7.3 introduces the application of public transportation smart card data in the determination of travel behavior and places of work and residence of commuters. Section 7.4 introduces the application of public transportation smart card data to the prediction of bus arrival times. Section 7.5 presents a public transportation big data analysis and

visualization system that was developed based on an open-source geographic information system (GIS) platform under the theoretical framework of *E*-Science, which is used to realize the application of big data in public transportation operation and planning.

7.2 PUBLIC TRANSPORTATION BIG DATA PREPROCESSING METHOD

The previous section briefly introduced types of data from the Beijing public transportation system. In the actual public transportation data mining process, researchers need to clean the original data after acquiring them. Data cleaning is the last step in the process of discovering and correcting identifiable errors in data files, including data consistency checking and handling of invalid and missing values. Because smart card and GPS data that are commonly involved in public transportation data mining are large in volume, they require special processing software for management, cleaning, and storage. This section primarily introduces cleaning and preprocessing methods for public transportation smart card and GPS data via case studies using data from the Beijing public transportation system.

7.2.1 Public Transportation Smart Card Data Cleaning

Mining and analysis of public transportation smart card data (Fig. 7.1) has two objectives: to assist the transit operators in making decisions about public transportation; and to provide a data basis for public transportation planning. Public transportation data preprocessing is a process of obtaining data that can be used for the subsequent data mining process by screening and correcting the data in the established database and statistical analysis tools. The following data-related problems should be considered when preprocessing public transportation smart card data [1, 2].

(1) Missing data

An occurrence of missing data will have an adverse impact on later-stage data processing. If key fields are missing, the analysis process may be impeded. Occurrences of missing public transportation smart card data are relatively uncommon. When certain data are missing, known data should be employed as a substitute for the missing data. If no data can be used as a substitute, inference may be made based on experience or known data.

(2) Data errors

Public transportation smart card data may contain some erroneous statistical data, e.g., time recorded as "24:30:00." To ensure that data analyses do not yield erroneous results, data validity should be examined and redundant data should be deleted. Classification and sorting methods should also be employed to discover the data that may cause noise in the analysis results.

	card ID	Type	Transaction date	transaction time	Route number	bus ID	Boarding stop	Exit stop	Driver ID	Conductor ID
1	1000751063284790	06	20150601	063611	24038	00047665	0	1	04003743	04005511
2	1000751021088624	08	20150601	101233	00849	00042305	23	10	10001507	10025646
3	1000751006774985	08	20150601	004336	01113	00008035	28	24	00001092	00001068
4	1000751084565249	08	20150601	002950	01113	00008035	24	12	00001092	00001068
5	1000751088078348	08	20150601	220031	01113	00008035	24	12	00001092	00001068
6	1000751069782551	08	20150601	003320	01113	00008035	24	11	00001092	00001068
7	1000751039790296	06	20150601	101839	00381	00023331	0	1	02000807	02010561
8	1000751039158930	06	20150601	102437	00381	00023331	0	1	02000807	02010561
9	1000751070149415	06	20150601	102408	00381	00023331	0	1	02000807	02010561
10	1000751001709850	06	20150601	120208	00381	00023331	0	1	02000807	02010561

FIG. 7.1 Example of public transportation smart card data.

(3) Redundant data

Redundant data contain repeated information. For example, repeatedly recorded data are the most simple and visual redundant data. The presence of redundant data increases data accuracy. When some data contain errors, information can be recovered using the redundant data. However, this redundant data will cause difficulties in later-stage data processing or erroneous analysis. For example, repeated smart card data recordings will cause an overestimation of ridership. Redundant data should be treated based on the specific objective of data analysis. In addition, space consumption due to data storage should also be considered.

(4) Data consistency

The source of public transportation system data involve multiple devices and departments. Data acquisition targets vary among different device manufacturers and departments, which generates a difference in data with the same meaning among different devices and departments. Data inconsistency may be an inconsistency in data accuracy, data units, or data storage formats, or an inconsistency in data definitions. Data unit and storage format problems can be addressed by standardization at the data preprocessing stage. However, data accuracy and definitions cannot be addressed after the data are collected. A fundamental solution to data accuracy and definition problems is to use a unified data specification to specify clearly data definitions, accuracy, and units.

(5) Stale data

As time progresses, some data that were acquired a long time ago may become stale and invalid. Stale data comprise a problem relative to the objective of data analysis. Stale data can be stored in a separate subdatabase.

(6) Removing useless fields

Datasheets contain numerous fields, some of which have no significance to the data analysis in this chapter. Therefore, these redundant fields can be removed to accelerate the data analysis process. The following fields have direct significance to the data analysis in this chapter: card identification number (card ID), transaction date, transaction time, public transportation smart card type, vehicle ID, route number, and record number. Thus, the fields amount received, card balance, city number, card issue number, monthly pass type, amount due, driver ID, and conductor ID can be deleted to simplify and accelerate the data analysis process.

7.2.2 GPS Data Cleaning

GPS data are primarily collected by on-board GPS systems. The GPS data format may be inconsistent with the actual data format that is required by the study. Therefore, the GPS data format needs to be converted prior to the data analysis process.

The GPS data of the Beijing public transportation system, as shown in Fig. 7.2, are selected as an example. These GPS data reflect some problems. Therefore, the following aspects should be considered when preprocessing these GPS data [3]:

(a) The route numbers obtained from parsing are incorrect. Correct route number and vehicle ID information should be determined based on the subscriber identity module (SIM) card information and the information in the vehicle information table.
(b) The up/down signs are incorrect. As a result, the traveling directions of the buses cannot be determined.
(c) The direction data should be multiplied by 10. For example, the direction that corresponds to 21 is 210 degrees.
(d) The coordinates should be translated by adding 0.006 to the longitude and adding 0.001 to the latitude to ensure the compatibility of the coordinate data with the GIS data of the bus routes.
(e) The number of information items recorded in each GPS file is limited. As a result, the GPS files are not completely compatible with the smart card data.
(f) The time recorded in each GPS file is Greenwich Mean Time (GMT). Therefore, 8 h should be added to the originally recorded time to convert it to Beijing Standard Time.

This section introduces preprocessing methods for public transportation smart card data and original GPS data. Processed data can be mined and analyzed. Starting in the following section, the application of public transportation big data in practical problems is introduced from the aspects of public transportation system planning, operation, management, and evaluation.

7.3 APPLICATION OF PUBLIC TRANSPORTATION DATA IN PLANNING

Based on the introductions in the previous two sections, public transportation smart card data contain rich travel information and represent the travel behavior of each passenger. Therefore, determining the spatial and temporal patterns of passengers and identifying their travel patterns can provide a basis for urban public transportation planning, public transportation stop layout optimization, and public transportation development, as well as data support for departments such as the public transportation group, traffic control department, and development and planning department in relevant ridership analysis, which can facilitate reasonable optimization of bus and rail transport. The previously mentioned analysis is important for identifying and analyzing the commuting behaviors of public transportation and the operational scheduling of public transportation for morning and evening peak hours. Therefore, this section explains how to mine and analyze the commuting behaviors of transit

	Up/down sign	time	Positioning state	latitude	longitude	speed	directon	route number
1	4	20150518080533	A	39.8660966666667	116.440315	0.1852	28	69300
2	4	20150518080537	A	39.8660766666667	116.440291666667	0.7408	30	69300
3	4	20150518080603	A	39.8659883333333	116.440183333333	0.926	0	69300
4	4	20150518080618	A	39.8659833333333	116.440161666667	0.5556	2	69300
5	4	20150518080621	A	39.8659883333333	116.440166666667	0.5556	2	69300
6	4	20150518080648	A	39.865985	116.440161666667	0.1852	3	69300
7	4	20150518080718	A	39.8659166666667	116.440108333333	0.3704	31	69300
8	4	20150518080722	A	39.8659316666667	116.44012	0.1852	30	69300
9	4	20150518080748	A	39.8660083333333	116.440178333333	0.1852	28	69300
10	4	20150518080818	A	39.866005	116.440163333333	0.1852	24	69300

FIG. 7.2 Example of original GPS data of the Beijing public transportation system.

passengers based on public transportation data via case studies using Beijing public transportation smart card data.

7.3.1 Extraction of the Commuting Characteristics of Public Transportation Passengers

The term "commute" refers to the travel of a person between his/her place of residence and place of work. Four aspects are considered when extracting commuting characteristics: the number of travel days and departure time from a temporal pattern perspective, and the commuter stops and commuter routes from a spatial pattern perspective. Methods for extracting commuting characteristics and calculating the commuting indices are as follows.

Step 1: Same departure times

Departure times are measured in half-hours. One day is divided into 48 periods and denoted by 0–47 (0 signifies that the departure time falls between 0:00 a.m. and 0:30 a.m.; 47 signifies that the departure time falls between 11:30 p.m. and 0:00 p.m.). The most frequent departure time for the home-to-work trip (T_h) and the most frequent departure time for the work-to-home trip (T_w) are extracted.

For example, for the holder of the card numbered 5389, $T_h = 12$ with a total of 9 occurrences of T_h, and $T_w = 18$ with a total of 10 occurrences of T_w. Thus, a total of $9 + 10 = 19$ occurrences of the most frequent departure time is observed.

Step 2: Number of travel days (N_{day})

N_{day} is the actual number of days when the commuter travels compared with the number of times that the card has been used.

For example, card ID 5389 has been used 50 times on 15 separate days in June. Therefore, $N_{day} = 15$.

Step 3: Same commuter stops

Stops are classified into two types: place-of-work stops and place-of-residence stops. The origin (O) and destination (D) of the first trip are considered the place of residence and place of work, respectively. The O and D of the last trip are considered to be the place of work and place of residence, respectively. The most frequent place-of-work stop (S_w) and the most frequent place-of-residence stop (S_h) are extracted.

For example, the information extracted from the usage of card ID 5389 indicates that the card has been used 25 times at the most frequent place-of-residence stop and 26 times at the most frequent place-of-work stop. Thus, the card has been used $25 + 26 = 51$ times at the most frequent stops.

Step 4: Same travel routes

For the same commuter routes, the most frequent routes of the first and last trips and the number of trips within a month are recorded.

Step 4.1: To address the problem in which the commuter takes different routes from the same O to the same D, the number of trips is supplemented based on the identified most frequent place of residence and most frequent place of work.
Step 4.2: The most frequent routes of the first and last trips are removed.
Step 4.3: The number of routes of the first trip with O as the most frequent place of residence and D as the most frequent place of work is determined. The number of routes of the last trip with O as the most frequent place of work and D as the most frequent place of residence is determined.
Step 4.4: The sum of the previously mentioned four types of routes is calculated.

For example, the information extracted from the usage of card ID 5389 indicates eight occurrences of the most frequent route for the home-to-work trip, six occurrences of the most frequent route for the work-to-home trip, two occurrences of a nonmost frequent commuter route for the home-to-work trip, and four occurrences of a nonmost frequent commuter route for the work-to-home trip. Thus, a total of $8 + 2 + 6 + 4 = 20$ occurrences of the four types of routes are observed. Table 7.1 lists an example of extraction of commuting indices.

The holder of card 32,664 is now taken as an example. This commuter often travels from his/her place of residence (stop 265) to his/her place of work (stop 723) by taking route 57,300 and then transferring to route 60,366 between 8:00 a.m. and 8:30 a.m. In the afternoon, this commuter travels from his/her place of work to his/her place of residence by taking route 60,366 and then transferring to route 00741.

7.3.2 Identification of Commuters and Estimation of Their Places of Work and Residence

Commuters are determined and identified using a rating method based on the four commuting indices (parameters) extracted in the previous section. The determination of commuters consists of two parts. The first part is to determine the baseline score for the rating process based on the iterative self-organizing data analysis technique (ISODATA) algorithm [4], which is employed to assist the determination of the commuter baseline. The baseline score is also used to classify commuters into different groups. For example, based on their scores, commuters are classified into absolute commuters, average commuters, and noncommuters. The second part is to rate each commuter based on the technique for order of preference by similarity to ideal solution (TOPSIS) algorithm [5] and calculate relevant scores based on the four indices for commuting characteristics, from which the travel behavior of each commuter is determined. Refer to [6] for the details of the algorithm.

Next, the algorithm is validated with actual data. Data of the Beijing public transportation system from June 1 to June 30, 2015 are employed as an example.

TABLE 7.1 Extraction of Commuting Characteristics

Card ID	S_h	S_w	R_h	R_w	T_h	T_w	N_{day}	N_{route}	N_{stop}	N_{time}
43,207	20,152	365	00986	00958	19	29	5	2	3	3
41,610	217	102	6–2	10–6	14	31	2	2	4	2
32,558	1533	5503	00359	00359	15	35	24	21	33	11
32,664	265	723	57,300–60,366	60,366–00741	16	36	15	10	19	9
86,147	10,485	20,295	00012	00012	19	25	3	5	4	2

Notes: S_h *is the most frequent place-of-residence stop.* S_w *is the most frequent place-of-work stop.* R_h *is the most frequent route for the home-to-work trip.* R_w *is the most frequent route for the work-to-home trip.* T_h *is the most frequent departure time of the home-to-work trip.* T_w *is the most frequent departure time of the work-to-home trip.* N_{day} *is the number of travel days.* N_{route} *is the sum of the numbers of occurrences of the most frequent routes.* N_{stop} *is the sum of the numbers of the similarity of stops.* N_{time} *is the sum of the numbers of occurrences of the most frequent departure times.*

The data consist of subway data (107,481,320 items), urban bus data (215,144,507 items), and suburban bus data (42,220,547 items). Commuters are rated based on four indices: the number of travel days, same commuter routes, same commuter stops, and same departure times.

The algorithm identifies a total of 4,706,010 commuters. The extracted most frequent place-of-residence stops are projected onto a GIS map (shown as a heat map in Fig. 7.3).

The heat map in Fig. 7.3 visually depicts the distribution of places of residence of Beijing public transportation commuters. The dark color (red color) indicates a relatively high commuter population density, and the light color (blue color) indicates a relatively low commuter population density. As demonstrated in the heat map, most of the places of residence are distributed along the subway lines.

Some typical places of residence can also be observed on the heat map, e.g., the places of residence around Shahe Station and the Life Science Park station along the Changping line in the north, Tiantongyuan Station and Huilongguan

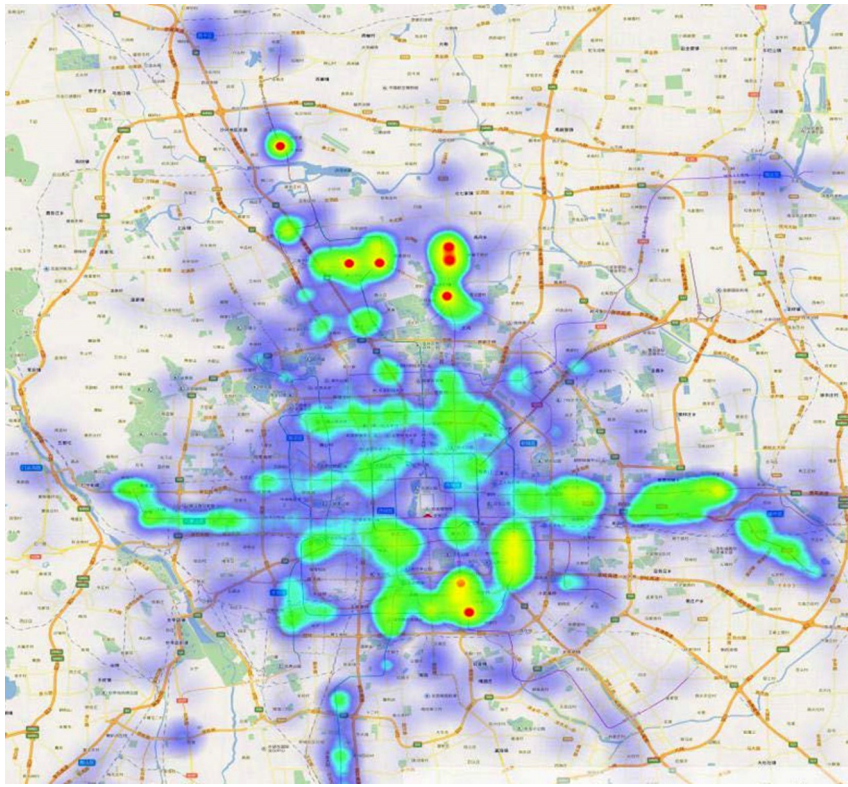


FIG. 7.3 Heat map of the places of residence of Beijing public transportation commuters in June 2015.

Station along Line 13, Puhuangyu Station and Liujiayao Station along Line 5 in the south, Shuangjing Station in the east and Caofang Station along Line 6. These places of residence are not only representative but also consistent with the actual situation. Fig. 7.4 displays the heat map of the extracted most frequent places of work.

The distribution of the places of work of Beijing public transportation commuters is visualized in the heat map in Fig. 7.4. Compared with the places of residence, the places of work in Beijing are concentrated and primarily distributed in the West 2nd Ring Road region, the East 2nd Ring Road region, the West 3rd Ring Road region, the Zhongguancun region, and the Xi'erqi region. The Central Business District has the highest place-of-work density, followed by Zhongguancun. Southern Beijing has a relatively low place-of-work distribution density, which is consistent with the actual situation.

To understand the distribution of commuters along the ring roads in Beijing, clustered stops with latitude and longitude coordinates are projected onto a GIS map. They are then classified into six groups based on the ring roads where they are located, as shown in Fig. 7.5: stops outside the 6th ring road; stops between the 6th ring road and the 5th ring road; stops between the 5th ring road and the

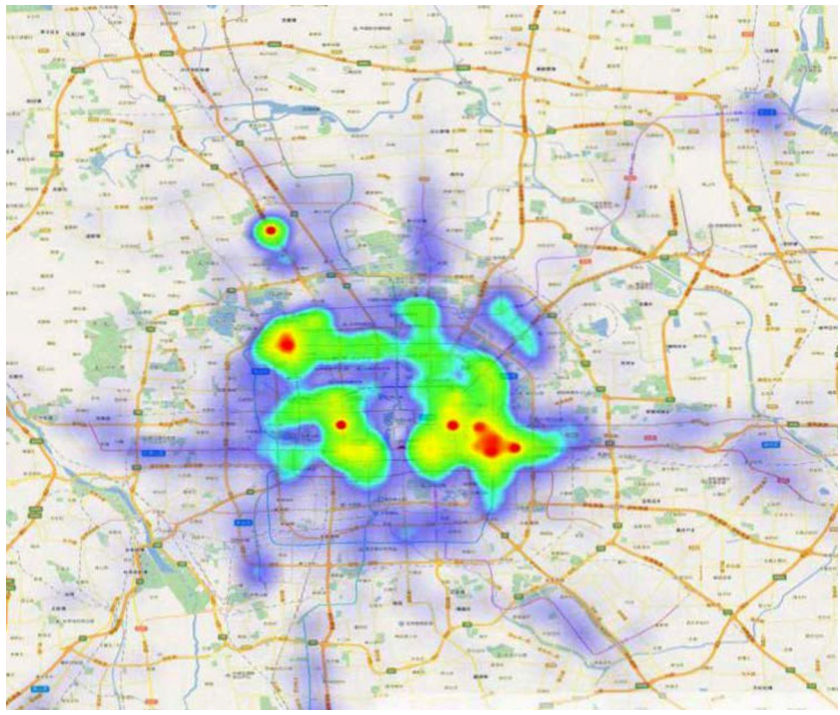


FIG. 7.4 Heat map of the places of work of Beijing public transport commuters in June 2015.

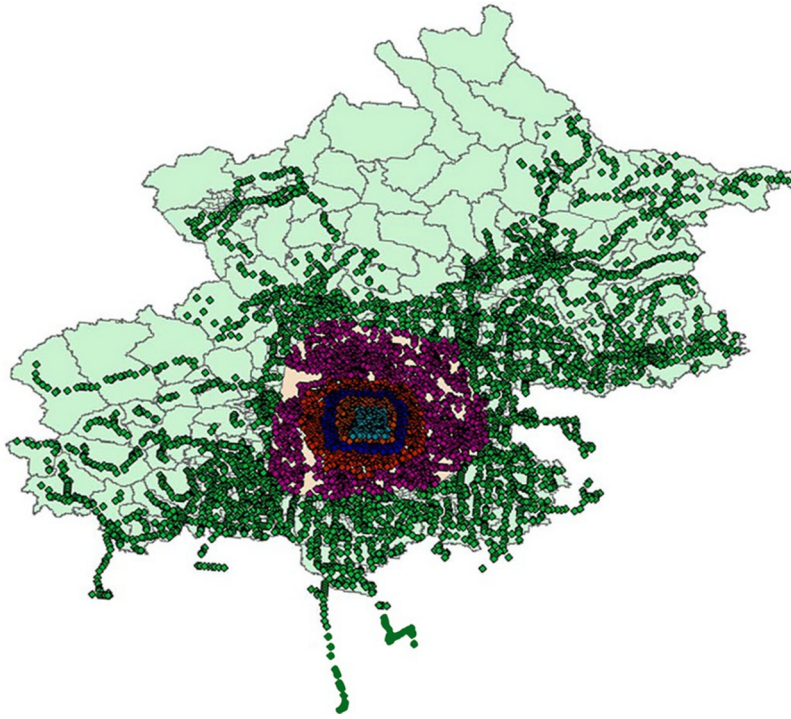


FIG. 7.5 Classification of stop IDs based on the ring roads where they are located.

4th ring road; stops between the 4th ring road and 3rd ring road; stops between the 3rd ring road and 2nd ring road; and stops within the 2nd ring road.

Based on the classification of stop IDs on the ring roads, the distributions of the places of residence and work of commuters can be analyzed. Table 7.2 lists the numbers of commuters at places of residence and work on each ring road and their percentage of the total.

As demonstrated in Table 7.2, most commuters have residences located outside the 3rd ring road, which accounts for approximately 83% of commuters. The 3rd ring road primarily contains busy commercial and administrative districts. Compared with other regions, the residence region between the 6th ring road and 5th ring roads has the largest number of commuters. The number of commuters with a residence outside the 5th ring road accounts for approximately 47% of the total number of commuters, which is attributed to lower housing prices or rents. Compared with the residences of commuters, most commuters work at places within the 6th ring road, which accounts for approximately 91% of all commuters. The number of commuters varies insignificantly between the work region between the 6th ring road and the 5th ring roads, the work region between the 5th ring road and the 4th ring road,

TABLE 7.2 Numbers of Commuters at Places of Residence and Work on Each Ring Road and Their Percentage of the Total

Ring Road	No. of Commuters at Resident Places	Percentage	No. of Commuters at Work Places	Percentage
Outside the 6th	576,471	12.42	391,572	8.43
6th–5th	1,638,422	35.31	821,710	17.68
5th–4th	783,918	16.90	725,858	15.62
4th–3rd	846,079	18.24	1,146,259	24.66
3rd–2nd	513,459	11.07	918,226	19.76
Within the 2nd	281,424	6.07	644,010	13.86

the work region between the 4th ring road and the 3rd ring road, the work region between the 3rd ring road and the 2nd ring road, and the work region within the 2nd ring road. In addition, most commuters (45% of all commuters) work in the region between the 4th ring road and the 2nd ring roads, where most of the commercial districts and places of work are distributed.

Based on the distribution of the places of work and residence of commuters along the ring roads, we can visually and comprehensively determine that the places of work are centrally distributed, whereas the places of residence are distributed along the subway lines. In addition, the places of work and residences of commuters are significantly unevenly distributed. These results can be used as data support for urban planning departments.

7.4 APPLICATION OF PUBLIC TRANSPORTATION DATA IN OPERATION AND MANAGEMENT

Informatization of an urban public transportation system can improve the effective utilization rate of public transportation vehicles to achieve dynamic scheduling and operation management of the public transportation system. Therefore, informatization of public transportation is important content for improving an urban public transportation system. For passengers, informatization can also provide tremendous convenience. By dynamically acquiring and analyzing the traveling information of public transportation vehicles and road and traffic information using advanced information and communication technologies and publishing the results via electronic public transportation stop boards, cell

phones or the Internet in real time, information, such as the levels of public transportation congestion, travel time of public transportation vehicles and transfers, can be provided to passengers. This section introduces the application of public transportation data in the prediction of bus arrival times, i.e., the prediction of passengers' waiting times and their confidence intervals. The range of bus arrival times is introduced to quantify the uncertainty of prediction results, which can help passengers become aware of the operational information of public transportation and enable them to make reasonable adjustments to their travel time.

7.4.1 Prediction Model for Public Transportation Bus Arrival Times

This section introduces a smart card data-based prediction method for the range of public transportation bus arrival (passengers' waiting) times. This method predicts bus arrival times and their confidence intervals based on a relevance vector machine (RVM).

The RVM, which was proposed by Tipping in 2001, is a learning machine that was constructed based on a sparse Bayesian framework [7]. It is a supervised machine-learning algorithm that is often employed to solve regression and classification problems [8].

Next, variables that affect bus arrival times are selected based on smart card data. The following key fields are extracted from the smart card data: the datasets of smart card IDs, bus IDs, transaction times, route numbers, stop IDs, arrival times, dates, number of boarding passengers, and number of alighting passengers.

The arrival time of a bus at a certain stop is irregular and affected by a number of factors, such as the level of traffic congestion and the number of boarding and alighting passengers at each stop. The level of traffic congestion directly affects the travel time of a bus between two stops. The number of boarding and alighting passengers can affect the dwell time that a bus waits at a stop. The arrival time of a bus can be predicted by predicting the bus headway at the target stop. Bus headway is a measurement of the difference between the arrival times of two consecutive buses at a certain stop. If the time headway between two buses at the previous stop is known, the arrival time of the back bus at the following stop can be determined by predicting its time headway at the following stop. Therefore, the headway of a bus at the previous stop is an important factor that affects its arrival time at the following stop.

Stops A and B are assumed to be two stops on a bus route, and bus i and bus $i + 1$ (the bus of the same route following bus i) are assumed to pass stops A and B successively. Stops A and B can be two adjacent stops or two stops that are relatively far apart. Stop B is the target stop for the bus headway prediction. Thus, stop A is the upstream stop. In this section, the following six variables are selected as factors that affect bus arrival times, which are used as the input features of the model:

(1) bus headway at the upstream stop A, h_A;
(2) number of passengers that board on bus i at the upstream stop A, $b_{i,A}$;
(3) number of passengers that alight from bus i at the upstream stop A, $a_{i,A}$;
(4) number of passengers that board on bus $i + 1$ at the upstream stop A, $b_{i+1,A}$;
(5) number of passengers that alight from bus $i + 1$ at the upstream stop A, $a_{i+1,A}$; and
(6) travel time of bus i between the upstream stop A and the target stop B, $T_{i,A}T_{i,B}$.

The number of boarding and alighting passengers can be calculated based on the total number of smart card transactions that occur at each stop. The travel time of a bus between two stops is equal to the difference between the arrival times of the bus at the two stops, as shown in Eq. (7.1). In this equation, $T_{i,j-1}$ represents the arrival time of bus i at stop $j - 1$, and $T_{i,j}$ represents the arrival time of bus i at stop j.

$$TT_{i,j-1} = T_{i,j} - T_{i,j-1} \tag{7.1}$$

Then, the bus headway at target stop B (h_B) is the output variable of the model.

7.4.2 Case Study

Based on the previously mentioned analysis, six factors affect bus arrival times. These six factors are used as the input features of the RVM model. Refer to [9] for the details of the algorithm.

A case study of the Chang 51 route of the Beijing automatic fare collection (AFC) system is performed. Data of the Chang 51 route from July 1 to October 31, 2012 are selected, of which the data of the first 3 months are used to train the RVM model and the data of the last month is used to test the RVM model. To verify the validity and effectiveness of the RVM algorithm, it is compared with a support vector machine (SVM) algorithm. Two evaluation indices—mean absolute percentage error (MAPE) and root-mean-square error (RMSE)—are established as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{h(i) - \hat{h}(i)}{\overline{h}}\right| \times 100\% \tag{7.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[h(i) - \hat{h}(i)\right]^2} \tag{7.3}$$

where $h(i)$ and $\hat{h}(i)$ are the real and predicted bus headway values at time i, respectively. Fig. 7.6 shows the comparison of the true values and the predicted values using the RVM and SVM algorithms. As demonstrated in Fig. 7.6, the predicted values that are obtained using the RVM algorithm are similar to the true values. As shown in Table 7.3, the MAPE and RMSE of the RVM

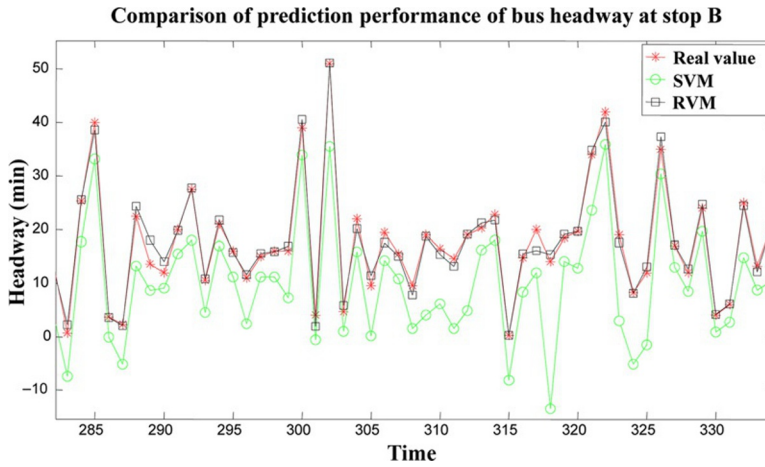


FIG. 7.6 Comparison of the true values and the predicted values that are obtained using the RVM and SVM algorithms.

TABLE 7.3 Errors of the RVM and SVM Algorithm

Algorithm	MAPE (%)	RMSE
RVM	14.63	3.5147
SVM	35.76	8.8054

algorithm are less than the MAPE and RMSE of the SVM algorithm, which demonstrates the validity of the RVM algorithm.

Fig. 7.7 shows the prediction results of the confidence interval of the bus headway based on the RVM algorithm (confidence level = 95%). As demonstrated in this figure, most of the real values fall within the confidence interval (RVM- and RVM+ represents the lower and upper bound of the confidence interval of the predicted values, respectively). The mean value of the actual bus headways is 18 min, whereas the width of the confidence interval is approximately 7 min, which indicates that passengers only need to arrive at the bus stop 7 min earlier. This interval will help improve passengers' travel experience. Thus, the effectiveness of the RVM algorithm is verified.

7.5 INTRODUCTION OF A PUBLIC TRANSPORTATION BIG DATA PLATFORM BASED ON *E*-SCIENCE

The previously mentioned analysis that relates to public transportation big data is important for improving public transportation service quality, optimizing the public transportation planning strategy, and motivating people to use public

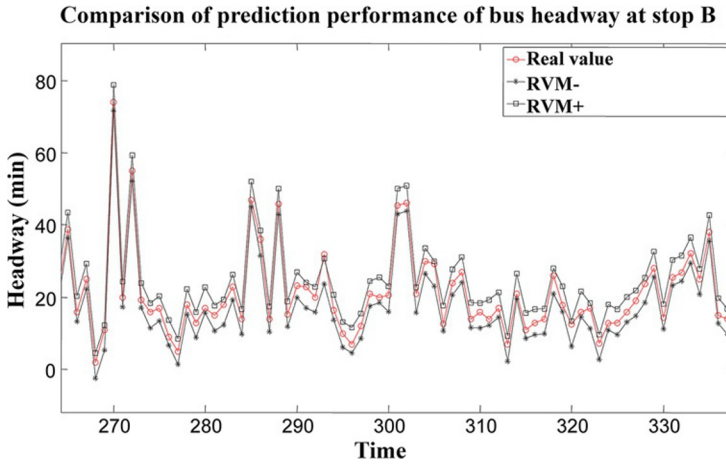


FIG. 7.7 Comparison of the confidence interval of the predicted values that are obtained using the RVM algorithm and the true values.

transportation. In addition, the previously mentioned analysis is crucial to the reasonable evaluation of the operating conditions of the public transportation network. An advanced public transportation network performance evaluation system can enable the public transportation department to make decisions and determine when, where, and how to provide better transport services. In addition, this system not only can facilitate public transportation departments to produce technical summaries and perform statistical analysis of the performance and operating conditions of public transportation networks, but also can make passengers aware of the operating conditions of urban transportation networks by publishing relevant information. Based on these reasons, an *E*-Science platform for public transportation is constructed in this section based on the theoretical framework of *E*-Science to evaluate the performance of public transportation networks.

E-Science was proposed in the United Kingdom as a brand new mode of scientific research that was established based on Internet technology and wide-area distributed high-performance computing environments to address the unprecedentedly complex problems in various disciplines and research fields, i.e., scientific research activities in support of information infrastructure.

Thus, this section primarily introduces a public transportation big data platform that is based on the theoretical framework of *E*-Science [10], which was established using a combination of spatial data (geospatial data of public transportation network) and public transportation data (public transportation smart card data)—an *E*-Science public transportation performance evaluation platform. This platform is an *E*-Science platform with integrated public transportation data sharing, visualization, modeling, and analysis functions. This platform can calculate public transportation performance indices, evaluate the performance of a public transportation network, and display the results to users.

7.5.1 Main Functions of the Public Transportation Big Data Platform

The Beijing AFC system and automated vehicle location (AVL) system data are the two main data sources for this platform. The Beijing AFC system entered service in the public transportation system on May 10, 2006. This platform provides various performance evaluation indices for public transportation [11]. Performance evaluation indices of various levels for public transportation are introduced as follows.

7.5.1.1 Network-Level Evaluation Index for Operating Speed

A minimum travel time-based transport route search algorithm that employs real-time traffic speed data is used to enable travelers to plan routes and formulate travel plans. By observing low-speed routes, public transit agencies can also identify congestion bottlenecks and study the spreading process of traffic congestion over networks. Then, public transit agencies can mitigate traffic congestion by relevant policies (e.g., transport route optimization). Smart card transaction data include time information (transaction time) and spatial information (public transportation routes and O and D information). Based on smart card transaction data, the travel time between two stops can be estimated. The travel time between two adjacent stops is composed of operating time and delay time. Delay time refers to the length of time that a bus waits at a certain stop to load and unload passengers, which is highly correlated to the number of passengers who wait at the stop. Therefore, when calculating the actual operating speed, the general traffic conditions and the delay time caused by passengers should be considered. Thus, the average speed of a bus between two stops can be calculated by dividing the network distance between the two stops by the travel time of the bus between the two stops. The network-level evaluation index for traffic speed can provide bus drivers with dynamic route information and public transportation departments with optimized public transportation networks and public transportation researchers with traffic congestion diagnoses.

7.5.1.2 Route-Level Reliability Evaluation Index for Public Transportation Travel Time

Travel time reliability represents the temporal stability of repeated trips. Providing travel time reliability information can help passengers plan their trips and reduce traffic congestion. In public transportation, public transportation travel time reliability affects the attractiveness and efficiency of transport services and is related to the on-time performance, which is indicative of customers' satisfaction and the stop-level headway variance. From the passengers' perspective, their travel time encompasses on-board travel time and waiting time. Route-level transport travel time reliability can be used to measure the variance of on-board travel time. Numerous effective methods are available for quantifying transport travel time reliability.

(1) 90th or 95th percentile travel time: this describes travel time during heavy traffic conditions.
(2) Buffer index: this calculates the time spent by a passenger in addition to the average travel time to ensure an on-time arrival. Additional time can be defined as the difference between the 95th percentile travel time and the average travel time. Thus, the buffer index is the ratio of the additional time to the average travel time.
(3) Scheduled time index: this is the ratio of the 95th percentile travel time to the travel time in smooth traffic conditions.

7.5.1.3 Stop-Level Ridership Evaluation Index

The stop-level ridership at each stop along a route refers to the number of boarding and alighting passengers at the respective stop. Ridership serves an important role in facilitating the public transit agencies to monitor transport services and attain economic benefits. Transport operators can identify the stops where with a large number of boarding passengers based on the stop-level ridership and adjust the transit schedules to improve public transportation service quality. Public transportation decision makers can evaluate the effectiveness of a new fare policy and respond to fare changes based on the public transportation ridership. The total number of boarding and alighting passengers determines the passenger transport demand of each route and affects the marketing and operational strategies of public transportation departments.

7.5.1.4 Stop-Level Headway Variance

Bus headway is a key factor that measures transport service reliability; it can be defined as the difference between the arrival times of two consecutive buses at a certain stop along the same route. Bus drivers and passengers prefer constant bus headways. A small bus headway will cause bus bunching, whereas a large bus headway will produce an increase in passengers' waiting time. Irregular headways will tarnish the attractiveness of public transportation. Because changes in traffic signals and the number of boarding and alighting passengers will cause changes in the bus headways at different stops, for transit schedules, stop-level headways must be analyzed to adjust the headways in the middle of the route.

7.5.2 Functions of the Public Transportation Big Data Platform

The four core components of this platform are as follows.

7.5.2.1 Public Transportation Network Speed Map

The travel speed of public transportation is calculated based on the identified passenger O-D pairs. Fig. 7.8 shows the traffic conditions of the Beijing public transportation network between 4:30 p.m. and 5:00 p.m. on July 31, 2010 (a workday).

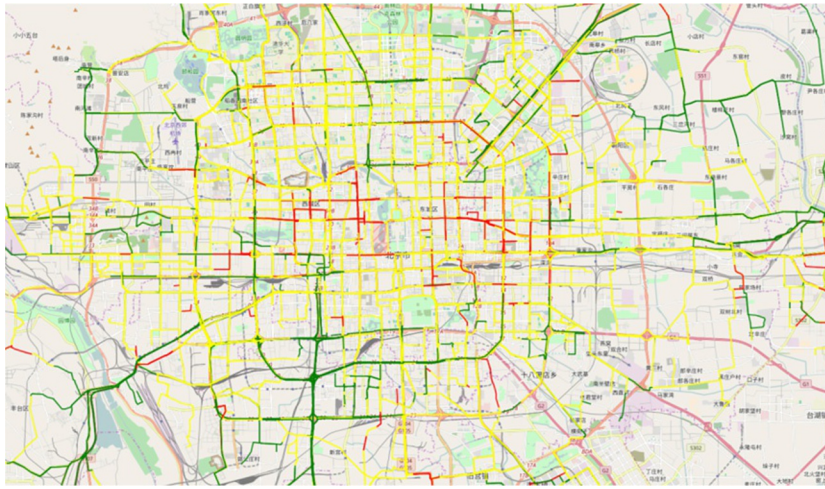


FIG. 7.8 Beijing public transportation network speed map. *(Map data sources:* Open Street Map.*)*

This platform provides an additional analysis function for checking the statistical data of the travel speed of public transportation within the network. A window that shows detailed traffic speed information about the entire network (e.g., average speed, variance, the 90th percentile speed, percentage of traffic sections that are congested/uncongested, composition of data sources for bus speed calculation (GPS and smart card data)) will pop up when the operating speed statistics button is clicked. Public transportation departments can use the color speed map and statistical analysis as an effective tool for identifying congested areas and make corresponding improvements in its public transportation services, e.g., opening high-speed roads or reducing headway for public transportation vehicles.

7.5.2.2 Public Transportation Ridership Analysis

Public transportation ridership analysis is important to public transportation departments. They strive to retain existing public transportation passengers and attract potential public transportation passengers. In addition, the spatial and temporal public transportation ridership shown on a map system can facilitate public transportation departments to perform a comparative analysis and evaluate changes in passengers' demand. As demonstrated in Fig. 7.9, the radius of each stop indicates the magnitude of the passenger load. The passenger load at each stop is defined as the difference between the number of boarding passengers and the number of alighting passengers.

7.5.2.3 Distribution of Bus Headways

Fig. 7.10 shows the bus headway of the No. 118 bus route on April 7, 2008. Estimation based on the smart card transaction data indicates a total of 116

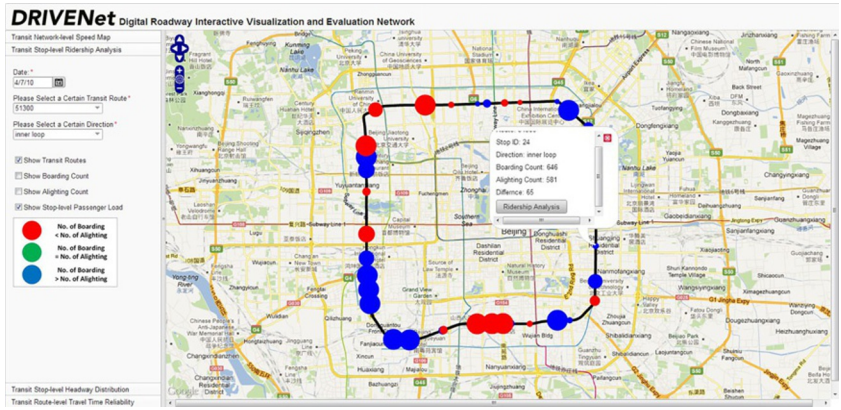


FIG. 7.9 Analysis of the ridership of route 51,300. *(Map data sources: Google and AutoNavi.)*

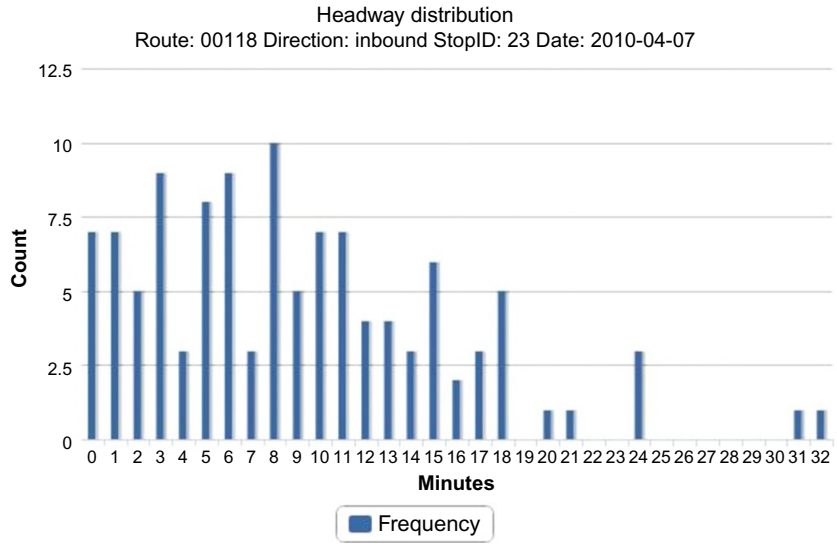
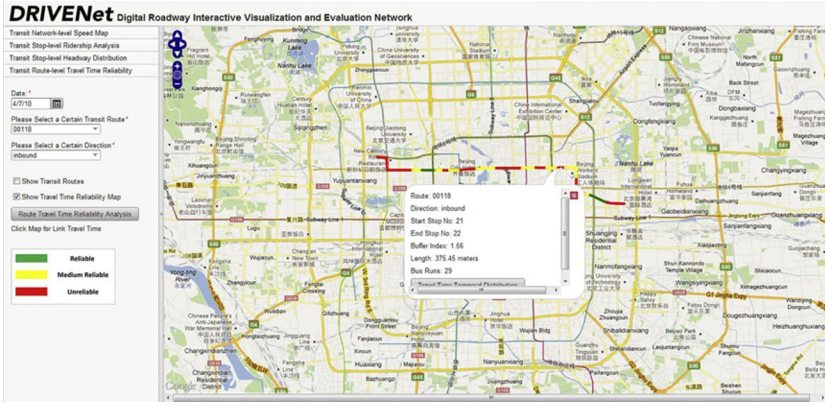


FIG. 7.10 A histogram of bus headways at a particular bus stop. *(Map data sources: Google and AutoNavi.)*

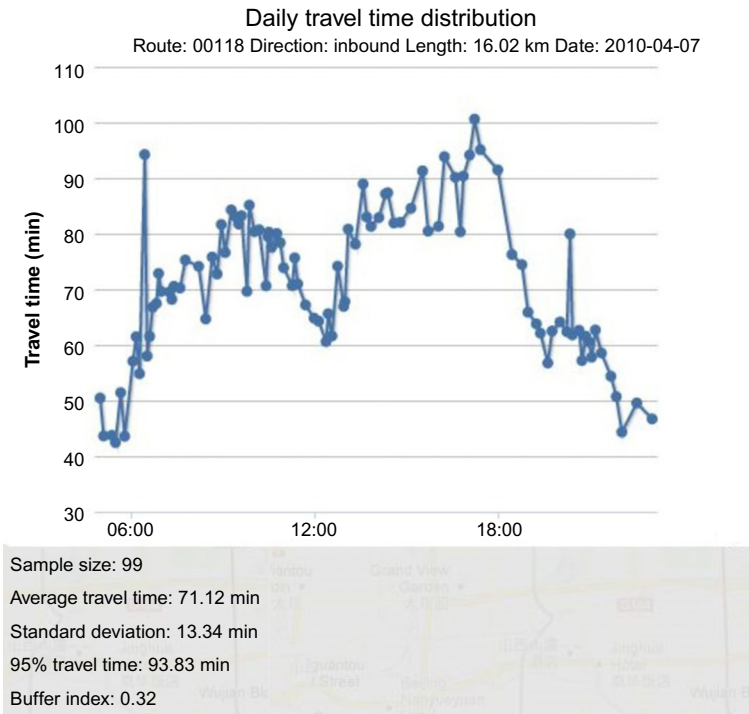
bus trips. The average bus headway and variance of bus headways at 23 stops can be calculated (9.47 min and 6.64 min, respectively). A frequency histogram is generated to display the distribution of headways.

7.5.2.4 Public Transportation Travel Time Reliability

Another function of this platform is to analyze public transportation travel time reliability based on a GIS map, as shown in Fig. 7.11A. The buffer time index is used to measure travel time reliability. The smaller the buffer time index is, the more reliable the public transportation route is. To analyze changes in the public



(A)



(B)

FIG. 7.11 (A) Spatial distribution of bus travel time reliability; (B) trend analysis of bus travel time. *(Map data sources: Google and AutoNavi.)*

transportation travel time, the travel time on each passing route/link can be calculated, as demonstrated in Fig. 7.11B. The buffer time index of the No. 118 bus route on April 7, 2010 is 0.32, which indicates that on-time arrivals with 95% can be ensured once bus passengers have waited 22.71 min.

7.6 CONCLUSIONS

In the big data era, a central issue is how to obtain correct, noise-free effective data from massive intelligent public transportation system data using efficient data processing methods, extract public transportation ridership, and vehicle information from these effective data. The extracted information then needs to be deeply mined to ensure that relevant analysis and prediction results are applicable to public transportation operation, planning, and management.

This chapter introduced the sources and types of data generated by public transportation systems against a big data background, as well as preprocessing methods for public transportation smart card and GPS data. Then, the application of big data in public transportation planning, operation, and management, and public transportation network evaluation was introduced. The validity and feasibility of each method proposed in this chapter was verified by a case study using Beijing public transportation system data. This chapter focused on transport data mining and analysis; the results can be used to provide public transportation information services to the public, reduce passengers' travel times on public transportation, improve the operational and management level of public transportation, and provide data support to the public transportation management department in their decision-making process.

ACKNOWLEDGMENT

This chapter is supported by the National Natural Science Foundation of China (61773036), Beijing Natural Science Foundation (9172011), and Young Elite Scientist Sponsorship Program by the China Association for Science and Technology (2016QNRC001).

REFERENCES

[1] M. Pelletier, M. Trépanier, C. Morency, Smart card data use in public transit: a literature review, Transp. Res. C Emerg. Technol. 19 (4) (2011) 557–568.

[2] S. Robinson, B. Narayanan, N. Toh, F. Pereira, Methods for pre-processing smartcard data to improve data quality, Transp. Res. C Emerg. Technol. 49 (2014) 43–58.

[3] X. Ma, C. Liu, J. Liu, F. Chen, H. Yu, Boarding stop inference based on transit IC card data (in Chinese), J. Transp. Syst. Eng. Inf. Technol. 15 (4) (2015) 78–84.

[4] G. Ball, J. Hall, A Novel Method of Data Analysis and Pattern Classification, Stanford Research Institute, 1965.

[5] C. Hwang, K. Yoon, Multiple attribute decision making: methods and applications, Lecture Notes Econ. Math. Syst. 375 (4) (1981) 59–191.

[6] X. Ma, C. Liu, H. Wen, Y. Wang, Y. Wu, Understanding commuting patterns using transit smart card data, J. Transp. Geogr. 58 (2017) 135–145.
[7] J. Lou, Y. Jiang, Q. Shen, Z. Shen, Z. Wang, R. Wang, Software reliability prediction via relevance vector regression, Neurocomputing 186 (C) (2016) 66–73.
[8] M. Tipping, Sparse Bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (3) (2001) 211–244.
[9] H. Yu, Z. Wu, D. Chen, X. Ma, Probabilistic prediction of bus headway using relevance vector machine regression, IEEE Trans. Intell Transp. Syst. 18 (7) (2017) 1772–1781.
[10] X. Ma, Y. Wu, Y. Wang, DRIVE net: an E-science of transportation platform for data sharing, visualization, modeling, and analysis, Transp. Res. Rec. 2215 (2011) 37–49.
[11] X. Ma, Y. Wang, Development of a data-driven platform for transit performance measures using smart card data and GPS data, J. Transp. Eng. 140 (12) (2014).