Smart Card Data Mining and Inference for Transit System Optimization and Performance Improvement

Xiaolei Ma

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:
Yinhai Wang, Chair
Edward D. McCormack
Bill Howe

Program Authorized to Offer Degree:
Department of Civil and Environmental Engineering

University of Washington

Abstract

Smart Card Data Mining and Inference for Transit System Optimization and Performance Improvement

Xiaolei Ma
Chair of the Supervisory Committee:
Professor Yinhai Wang
Department of Civil and Environmental Engineering

The United States energy information administration states that more that 50% of commuters drive their own cars to work. This implies that traffic congestion can be mitigated if public transit service can take a larger share of commuting trips. However, a commuter's choice depends on the utility associated with each available mode. Transit service must be improved to increase its utility and therefore attract more riders.

To improve customer satisfaction and reduce operation costs, transit authorities have been striving to monitor their transit service quality and identify the key factors to attract the transit riders. Traditional manual data collection methods are unable to satisfy the transit system optimization and performance measurement requirement due to their expensive and labor-intensive nature. The recent advent of passive data collection techniques (e.g. Automated Fare Collection and Automated Vehicle Location) has shifted a data-poor environment to a data-rich environment, and offered opportunities for transit agencies to conduct comprehensive transit system performance measures. Although it is possible to

collect highly valuable information from ubiquitous transit data, data usability and accessibility are still difficult to improve due to the following reasons: (1) most Automatic Fare Collection (AFC) systems are not designed for transit performance monitoring, hence additional passenger trip information cannot be directly retrieved. (2) Each passive data collection method has its intractable disadvantages, and requires additional domain knowledge to process. Interoperating and mining heterogeneous datasets would enhance both the depth and breadth of transit-related studies. (3) The amount of data involved is increasingly growing, and traditional data processing applications might not be suitable to handle in an efficient fashion. Such data barriers hinder the development of a large-scale transit performance monitoring system.

This study attempts to fill these research gaps by developing a series of data mining algorithms for transit rider's origin and destination information extraction with transit Smart Card (SC) data. The primary data source of this study comes from the AFC system in Beijing, where a passenger's boarding stop (origin) and alighting stop (destination) on a flat-rate bus are not recorded on the check-in and check-out scan. A Markov chain based Bayesian decision tree algorithm is proposed to mine the passengers' origin information using SC data. In addition, this study further proposes an integrated data mining procedure that models the travel patterns and regularities of transit riders. This procedure is able to incorporate transit riders' trip chains based on their temporal and spatial characteristics, and capture their historical travel patterns in an efficient manner. Then, on the basis of the identified travel patterns, the individual-level destination can be estimated with transfer analysis through a multi-day observation. Finally, to remove data accessibility barriers, facilitate data sharing

and visualization, and conduct online data analysis for transit performance measures, an e-science of transportation platform entitled TransitNet is developed. TransitNet enables the connections and interoperability among the heterogeneous transit data sets including SC data, GPS data and Geographic Information System (GIS) data. This platform not only serves as a data-rich visualization platform to monitor transit network performance for planning and operations, it also intends to take advantage of e-science developments for data-driven transportation research and applications.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to present my enormous appreciation to my PhD advisor and doctoral committee chair, Professor Yinhai Wang. His enthusiasm and diligence for research encourages me to enter the domain of transportation engineering. His vision and thoughtful ideas always inspire me to conduct in-depth scientific discovery. Without his kind guidance and constructive comments on my research during my five year PhD study, it would be difficult for me to complete my PhD dissertation.

I also want to thank Professor Ed McCormack, Professor Jenq-Neng Hwang and Professor Bill Howe to offer me valuable advice for my dissertation. Especially I would express my sincere gratitude to Professor Ed McCormack, who also served as my advisor during my master study. I really benefited a lot from his consistent support and supervision.

I would like to express my appreciation to Beijing Transportation Research Center (BTRC), where I spent one year in 2010 conducting Beijing transit data mining project. BTRC sets up a solid data foundation for my dissertation, and provides multiple valuable datasets including Beijing smart card transactions, GPS and GIS data. Specially, I want to thank Feng Chen and Jianfeng Liu, who frequently discussed with me about my research progress. Their feedbacks are incorporated to improve my algorithms. I also want to thank Wenhua Pu, who is one of the most hard-working software engineers that I have ever seen. He assisted me implementing my algorithms in a large-scale network.

I am also very grateful to have several STAR lab fellows to support and help me in the past five years. Particularly, I would like to thank Yao-Jan Wu, Runze Yu, Guohui Zhang, Matthew Dunlap and Sa Xiao. Yao-Jan and Guohui provided me significant suggestions and guidance at the beginning of my graduate study, and Runze is a great co-worker during my PhD study at UW, and we experienced all the bitterness and happiness. Matt spent lots of his time proof-reading my dissertation, and provided me valuable suggestions. Sa is a very talented master student, and she helped improve my second version of DRIVE Net system. I benefited from her implementation to develop TransitNet system for my PhD dissertation.

Last but not least, I would show my greatest gratitude to my family. Without their encouragements and supports, I would not be able to accomplish my PhD.

DEDICATION

To my family.

Chapter 1 Introduction

1.1 Problem Statement

Approximately 76% of those living in the United States chose privately owned vehicles for their commute to work in 2000 (ICF Consulting, 2003) and data collected for the 2009 American Community Survey indicate that 79.5% drive alone when commuting (McKenzie and Rapino, 2011). This pattern is now becoming apparent in developing countries such as China, where many rely on privately owned vehicles to commute. In 2010, for example, more than 34% of Beijing residents chose cars as their primary travel mode while only 28.2% chose transit (Beijing Transportation Research Center, 2011).

Public transit has been considered as an effective countermeasure to reduce congestion, air pollution, and energy consumption (Federal Highway Administration, 2002). According to the 2005 urban mobility report conducted by the Texas Transportation Institute (2005), travel delay in 2003 would have increased by 27 percent without public transit, and in the most congested metropolitan cites of U.S., public transit services have saved more than 1.1 billion hours of travel time. Moreover, public transit can help enhance business, and reduce city sprawl through transit oriented development (TDO). During certain emergency scenarios, public transit can even act as a safe and efficient transportation mode for evacuation (Federal Highway Administration, 2002). Based on the aforementioned reasons, it is of critical importance to improve the efficiency of public transit system, and promote more roadway users to utilize public transit. To fulfill these objectives, transit agencies need to understand the areas where further improvements can be made, and whether community goals are being met, etc. A well-developed performance measure system will help facilitate decision making for transit agencies. Transit agencies can evaluate the transit ridership trends with fare policy changes and identify where and when better transit service should be provided. In addition, transit agencies are also required to summarize transit performance statistics for reporting to either the National Transit Database (Kittelson & Associates et al., 2003), or the general public who are interested knowing how well transit service is being provided. Nevertheless, developing a set of structured performance measures often requires a large amount of data and the corresponding domain knowledge to process and analyze the data. These obstacles create challenges for transit agencies that must spend time and effort undertaking.

Traditionally, transit agencies heavily rely on manual data collection methods to gather transit operation and planning data (Ma et al., 2012). However, traditional data collection methods (e.g. travel diary, survey, etc.) are fairly costly and difficult to implement at a multiday level due to their low response rate and accuracy. Transit agencies have spent tremendous manpower and resource undertaking manual data collections, and consumed a significant amount of energy and time to post-process the raw data. With advances in information technologies in intelligent transportation systems (ITS), the availability of public transit data has been increasing in the past decades, which has gradually shifted the public transit system into a data-rich environment.

The Automatic Fare Collection (AFC) system and the Automatic Vehicle Track (AVL) system are two common passive data collection methods. To complete financial transactions, the AFC system, also known as the Smart Card system, records and processes the fare related information using either a contactless or a contact card (Chu, 2010). There exist two typical types of AFC systems: entry-only AFC system and distance-based AFC system. In the entry-only AFC system, passengers are only required to swipe their smart cards over the card reader during boarding. For the distance-based AFC system, passengers need to check in and check out during both their boarding and alighting procedures

AVL and AFC technologies hold substantial promise for transit performance analysis and management at a relative low cost. However, historically, both AVL and AFC data have not been used to their full potentials. Many AVL and AFC systems do not archive data in a readily utilized manner (Furth, 2006). AFC system is initially designed to reduce workloads of tedious manual fare collections, not for transit operation and planning purposes, and thereby, certain critical information, such as specific spatial location for each transaction, may not be directly captured. AVL system tracks transit vehicles' geospatial locations by Global Positioning System (GPS) at either a constant or varying time interval. The accuracy of GPS occasionally suffers from signal loss due to tall building obstructions in the urban area (Ma et al, 2011). Both AFC and AVL systems have their inherent drawbacks in monitoring transit system performance, and require analytical approaches to eliminate erroneous data, remedy missing values, and mine unseen and indirect information.

Conveying these processed AVL and AFC data in an understandable and flexible manner will be particularly desirable for transit agencies. In the past years, several home-grown software tools have been developed to facilitate transit agencies to generate key transit performance indicators; however, these tools lack the flexibility to adapt to their particular needs. Because of different data formats and software environments, stand-alone and expensive commercial software tools may be inaccessible to most agencies' staff analysts. In addition, the amount of transit data involved is intensively increasing and is challenging to manage using traditional data processing applications. On the other hand, developing one's own transit data processing and analytical tools are beyond most transit agencies' capabilities, and requires additional resources to handle. Few efforts have been made to propose an explicit architecture and framework for interactive online transit performance analysis with consideration of such heterogeneous data. Such a framework will greatly improve transit data usability and accessibility for transit agencies.

1.2 Research Background

1.1.1 Potential Use of Passive Data in Public Transit

Transit passenger Origin-Destination (OD) data are crucial for transit system planning and route optimization (Li, 2009). A transit rider OD pair can be potentially extracted from the SC transaction database. However, this is not a straightforward task. Two major challenges must be addressed to obtain good quality OD data. The challenges originate from the design of the SC scan system for the flat-rate buses. Since passengers pay a fixed rate to the flat-rate buses, only check-in scan is considered necessary in the SC scan system design (Zhao, et al., 2007). Compared to the distance-based fare bus riders, flat-rate bus users do not have check-out records. This creates the first challenge in OD extraction: where does a passenger get off a flat-rate bus? Furthermore, the scan system does not save the location and direction information of the check-in scans and this creates the second challenge: where does a passenger get on a flat-rate bus?

The two challenges induce two very interesting research topics: (1) how to identify the transit stop ID for a check-in scan, and (2) at which transit stop does the passenger get off the flat-rate bus? Given the fixed route of transit vehicles, known distance between stops, and transaction records stored in the database, such as smart card ID, route number, driver ID, transaction time, remaining balance, transaction amount, etc., it is not impossible to estimate a flat-rate bus user’s check-in and check-out stops through data mining and data fusion techniques. However, the accuracy of the extracted OD data depends largely on the quality of the data processing algorithms (Zhang, 2002).

To improve transit services and encourage more people to use public transit, transit agencies have been striving to identify the key factors that attract transit riders through studying their travel patterns. With a better understanding of the travel patterns of transit riders, transit authorities will be able to evaluate their current services to reveal how best to adjust their marketing strategies to encourage higher usage (Boyle et al., 2000). For example, knowing why some riders are especially loyal to transit can help transit agencies to determine where and when they should provide discounts to retain these loyal transit riders and potentially attract new riders (Trépanier et al., 2012). Based on identified travel patterns and transit usage regularities, transit authorities are able to evaluate the most cost-effective fare packages for transit riders, understand how transit riders' behaviors are likely to change in response to a new fare structure, and thus select a fare policy that achieves the optimum balance between enhancing the attractiveness of the transit system and maximizing fare revenue (Taylor and Jones, 2012).

In addition, transit planners and researches can also utilize individual travel-behavior data for activity-based trip modeling and transit travel demand analyses. For public agencies, information on the travel patterns for individual transit riders can also be utilized to quantify the effectiveness of transit-oriented development (TOD) (Dill, 2008). In particular, personal travel behavior data can reveal how TOD residents change their daily commuting behaviors and how transit use varies spatially and temporally. However, acquiring individual transit travel pattern is challenging (Tirachini, 2012). Traditional transit travel pattern analysis largely relies on rider satisfaction surveys or travel diaries (Chu and Chapleau, 2010), which is very costly and difficult to implement at a multiday level due to the low response rate and accuracy. The use of, smart card data to track passengers' long term travel activities and patterns, such as the number of typical daily trip chains, common boarding/alighting stops and trip start/end times, offers a far more convenient and efficient data source. Smart card data records both temporal and spatial information for each rider, making it feasible to conduct individual travel pattern analysis through longitudinal analyses (Chu, 2010).

Monitoring the performance of public transit system is a stepping stone toward the successful and proactive transit management (Bertini and El-Geneidy, 2003). In past decades, comprehensively evaluating the transit system has been considered difficult due to limited data sets from conventional collection techniques, such as surveys, questionnaires or interviews. Thanks to the repaid deployment of ITS sensors, collecting a wealth of good quality transit data at a relatively low cost is becoming more and more viable. Using these widely available data, transit agencies is able to evaluate their current transit service, better understand the ridership, and ultimately increase the attractiveness of public transit. For instance, transit agencies can examine the stop-level ridership, check the headway deviation, and consequently optimize transit routes to improve service reliability (Feng et al, 2011). Through visualization and GIS technologies, an interactive web-based transit information system would be of particular interest to transit rides to schedule their trip itinerary for time savings, and also reduces their waiting time by providing the real-time transit arrival information (Ferris, 2011; Sun et al, 2011). For transportation researchers, such a data-rich visualization platform will revolutionize the traditional transportation study from the mathematic-equation driven scope to the data-driven perspective (Peng and Huang, 2000). For instance, large-scale transit route optimization problems could be solved in an efficient and effective fashion by analyzing the network-level transit travel time data. Similarly, stops with the heavy ridership could be intuitively identified by observing the stop-level number of boarding passengers. Transportation researchers can also consolidate the development of transit assignment models by observing individual-level transit rider’s behavior (Hamdoucha et al, 2011). Establishing such an online transit performance measures system is beneficial for transportation practitioners, transportation researchers and transit riders.

1.1.2 Data Sources

Data from AFC system and AVL system are the two primary sources in this study. Beijing Transit Incorporated began to issue smart cards in May 10, 2006. The smart card can be used in both the Beijing bus and subway systems. Due to discounted fares (up to 60% off) provided by the smart card, more than 90% of the transit riders paid for their transit trips with their smart cards in 2010 (Beijing Transportation Research Center, 2010). Two types of AFC systems exist in Beijing transit: flat fare and distance-based fare. Transit riders pay at a fixed rate for those flat fare buses when entering by tapping their smart cards on the card reader. Thus, only check-in scans are necessary. For the distance-based AFC system, transit riders need to swipe their smart cards during both check-in and check-out processes. Transit riders need to hold their smart cards near the card reader device to complete transactions when entering or exiting buses. Smart cards can be used in the Beijing subway system as well, where passengers need to tap their smart card on top of fare gates during entering and existing subway stations. Both boarding and alighting information (time and location) are recorded by the fare gates. Although transit smart card exhibits its superiority in its convenience and efficiency, there are still the following issues to prevent transit agencies from fully taking advantage of smart cards for operational purposes:

- Passenger boarding and alighting information missing

Due to a design deficiency in the smart card scan system, the AFC system on flat fare buses does not save any boarding location information, whereas for distance-based fares, the AFC system stores boarding and alighting locations, but stores no boarding time information. Key information stored in the database therefore includes smart card ID, route number, driver ID, transaction time, remaining balance, transaction amount, boarding stop (for distance-based fare buses only), and alighting stop (for distance-based fare buses only).

- Massive data sets

More than 16 million smart card transactions data are generated per day. Among these transactions, 52% are from flat-rate bus riders. These smart card transactions are scattered in a large-scale transit network with 52386 links and 43432 nodes as presented in figure 1-1:

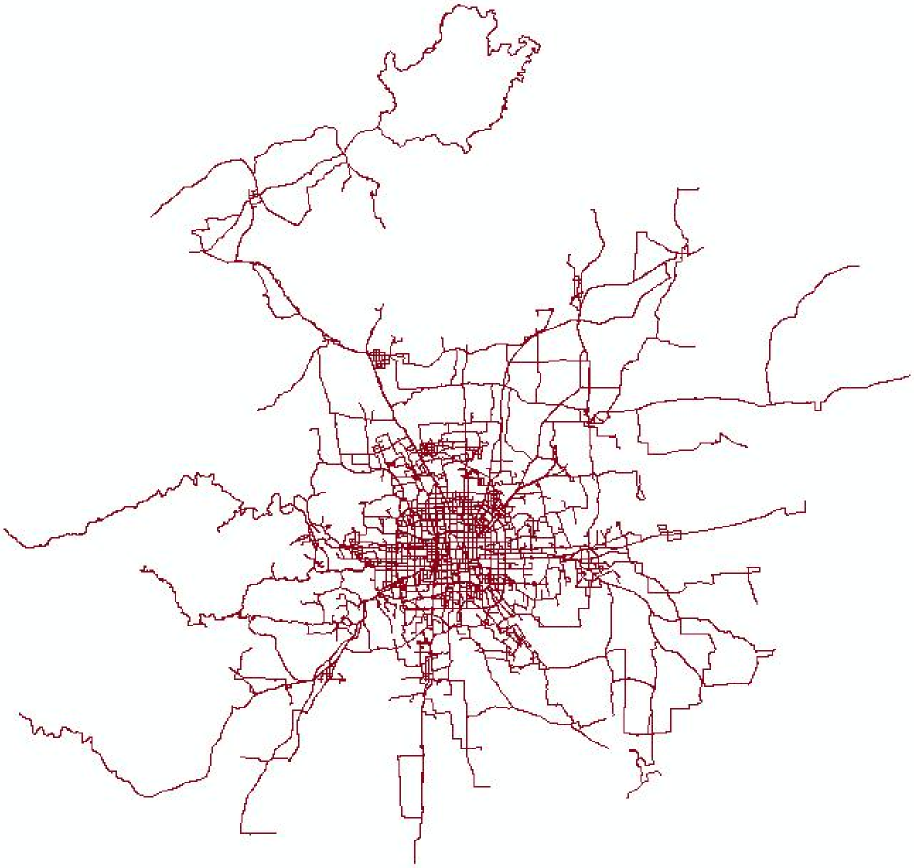


Figure 1-1 Beijing Transit Network in GIS

- Limited external data with poor quality

Only approximate 50% of transit vehicles in Beijing are equipped with GPS devices for tracking. GPS data are periodically sent to the central server at a pre-determined interval of 30 seconds. However, the collected GPS data suffer from two major data quality issues: (1) vehicle direction information is missing, and (2) GPS point fluctuation (Lou, et al., 2009). Map matching algorithms are needed to align the inaccurate GPS spatial records onto the road network. In addition, most of transit routes are not designed to have fixed schedules because of high ridership demands, and only certain routes with a long distance or headway follow schedules at each stop (Chen et al, 2009). These characteristics of the Beijing AFC and AVL systems create more challenges to process and mine useful information.

It is noteworthy that the AFC system used in Beijing is not a unique case. Most cities in China also employ the similar AFC system where passengers' origin information is absent, such as Chongqing (Gao and Wu, 2011), Nanning (Chen, 2009), Kunming (Zhou et al., 2007). Even in other developing countries, such as Brazil, the AFC system doesn't record any boarding location information as well (Farzin, 2008). Therefore, a solution for passenger boarding and alighting information extraction is beneficial to those transit agencies with imperfect SC data internationally.

1.3 Research Objectives

The ultimate objective of this study is to establish an e-Science platform to modeling, analyzing, and visualizing public transit passive data, and provides a solid, expandable foundation for road users, researchers, and decision makers. The key objectives of this study include:

- Develop a statistical model to infer passenger origin information using smart card data in an efficient and effective fashion;

- Develop a comprehensive data-mining procedure to extract each transit rider’s travel pattern and regularity information from large smart card datasets with incomplete information;
- Develop an individual-level passenger destination estimation algorithm considering transit riders’ travel behaviors;
- Develop a computational engine to mine passenger origin and destination based on smart card data;
- Identify several performance measures to evaluate public transit system
- Develop an e-Science-based transit spatial information platform for visualization and analysis

1.4 Study Scope

Automatic fare collection (AFC) systems contain rich spatial and temporal information obtained from through contactless smart cards with unique IDs, which significantly reduce the manpower required to collect transit passenger OD data. However, most AFC systems are not designed for OD data collection. Especially for AFC systems in the majority of developing countries, passengers’ boarding information is missing. Hence further data processing and analysis are necessary for passenger information extraction. This paper presents a Bayesian decision tree based statistical approach to infer the passenger origin from the imperfect SC transaction data, which is the first step for the transit OD estimation technique. In addition, smart card data collected by the AFC system would lack certain trip related information that affects data processing performance. To deal with this data issue, this paper proposes a robust and comprehensive data-mining procedure to extract individual transit rider's travel pattern and regularity using a large dataset with incomplete information. Then, a transit rider's behavioral information (i.e. travel patter and regularity) is incorporated into individual level passenger destination estimation, and the algorithm accuracy is improved. Finally, with the inferred individual level passenger OD data and processed trip information, a series of comprehensive performance indicators can be further developed to quantify the quality of public transit system. To visualize these performance measures in an understandable fashion for both transit agencies and transportation researchers, an e-science geospatial platform called TransitNet is developed to facilitate both transit system operational optimization and planning. The overall methodological framework is outlined in Figure 1-2.

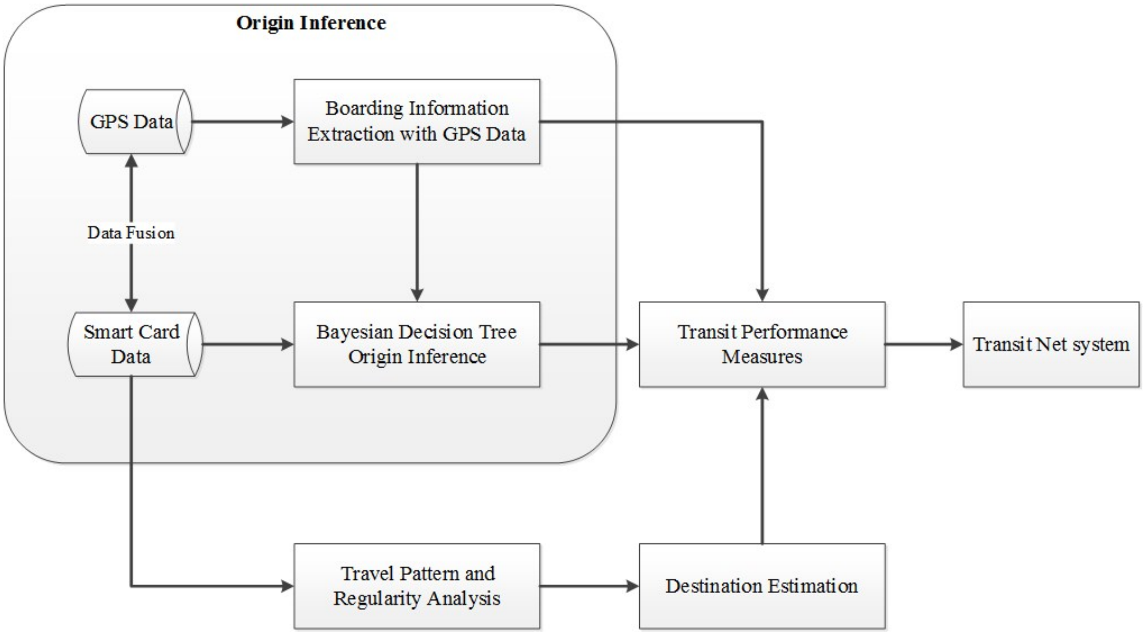

Origin Inference
GPS Data
Boarding Information Extraction with GPS Data
Data Fusion
Smart Card Data
Bayesian Decision Tree Origin Inference
Transit Performance Measures
Transit Net system
Travel Pattern and Regularity Analysis
Destination Estimation

Figure 1-2 Methodological Framework

1.5 Expected Benefits

In a short term run, the study provides insights on transit passive data retrieval, processing, storing and visualization. In particular, this study proposes a novel statistical approach to mine the individual-level passenger origin using incomplete and imperfect smart card datasets. Moreover, travel behavioral information of each transit rider can be inferred in an efficient and effective fashion. The understanding of travel behaviors allows transportation researchers to strengthen transit planning procedures, including activity-based trip modeling and dynamic transit assignment analysis. For the transit agencies, they can utilize the transit riders' travel patterns to identify the transit travel demand, adjust their marking strategies, and further improve the transit system service. On the basis of the mined individual-level travel pattern and travel regularity, a more robust passenger alighting stop algorithm is developed. By linking each passenger's origin and destination, transit agencies and transportation researchers can better conduct trip-related analyses, including: transfer activity detection, trip purpose identification, and stop-level or route-level travel demand forecasting, etc. Combined with other types of data (e.g. GPS data and GIS data), a visualization platform can be established to monitor and evaluate the transit system performance, assist transit agencies to optimize the transit network and ultimately increase the attractiveness of public transit. This will benefit reducing congestions and alleviating traffic-related pollution.

In the long run, this study lays a solid foundation to support future endeavors in the e-science area of transportation. The amount of data generated from people's daily lives has been growing rapidly. This is especially true in the transportation domain. With advances in data-collection technologies and their deployment in intelligent transportation systems (ITS), the availability and accessibility of transportation data have increased tremendously in the past decades. The size of data is so large that traditional data management tools cannot process it within a tolerable time frame. These large quantities of data are called “Big Data” (Zikopoulos, et al., 2011). “Big Data” initiative brings up both challenges and opportunities, and revolutionizes a variety of domains ranging from astronomy to bioengineering; however, the transportation community has shown a slow progress to accept this concept. Big data requires novel approaches to process the huge amount of data efficiently. Data mining and visualization techniques are two of the suggested suitable tools by McKinsey (Manyika et al., 2011). This study attempts to utilize these two weapons to bridge the research gap between public transit and the big data concept.

1.6 Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter 2 reviews several previous research efforts on transit smart card data applications, including passenger origin and destination inference approaches, individual-level passenger behavior analysis and mining, and existing transit performance measures programs. In addition, this chapter envisions the future of data-driven intelligent transportation system with a specific emphasis towards how “big data” can revolutionize the conventional mathematical-equation driven transportation studies.

Chapter 3 focuses on passenger origin information estimation using both GPS and smart card data. A data fusion method to integrate roadway geospatial data, transit vehicle GPS records and smart card transactions is firstly proposed to infer each passenger's boarding location. Moreover, a Bayesian decision tree algorithm is presented to estimate each passenger's boarding stop when GPS data are unavailable. Considering the expensive computational burden of decision tree algorithms, Markov-chain property is taken into account to reduce the algorithm complexity. Smart card transaction data and GPS data from the Beijing transit system are used to test and verify the proposed algorithms.

On a basis of the inferred individual passenger origin information, Chapter 4 further extracts each passenger's travel behavior information through multi-day observations. A robust and comprehensive data mining procedure is proposed to generate each individual's spatial/temporal travel pattern and travel regularity. The proposed data mining procedure is further optimized using the rough set theory to extract association rules from massive smart card transactions. The rough-set-based algorithm is compared with other prevailing classification algorithms. The results indicate that the proposed algorithm outperforms other prevailing data-mining algorithms in terms of accuracy and efficiency.

In Chapter 5, individual-level passenger alighting location can be estimated by the following three criteria: (1) Consider each passenger's transfer activity (2) Consider each passenger's daily trip characteristics (3) Integrate each passenger's historical travel pattern and travel regularity. The inferred individual passenger destinations are compared with ground-truth data from distance-based buses, where each passenger's boarding stop and alighting stop are recorded. The results imply that the proposed passenger alighting location inference algorithm can achieve a fairly high accuracy.

With individual passenger origins and destinations, transit performance measures are further developed in Chapter 6. Multiple-scale performance indicators are calculated based on both processed smart card data and GPS data. The key performance indices include network level travel speed, stop-level passenger ridership (i.e. number of boarding and alighting passengers), stop-level transit vehicle headway, and segment-level travel time reliability. To disseminate and convey these transit performance statistics in an efficient and effective fashion, an eScience platform for sharing, visualizing, modeling, and analyzing transit-related data are developed. This platform is named as TransitNet, and it not only serves as a data visualization and archival system, but also enables connections and interoperability among heterogeneous data sets including smart card data, GPS data and GIS data. This prototype gains significant insights from the conventional transportation research to more powerful data-driven solutions.

Finally, Chapter 7 concludes the research effort in this dissertation and envisions further research directions.

Chapter 2 State of the Art

2.1 Transit Origin and Destination

Many OD matrix inference approaches have been investigated over the past years. Researches on Metropolitan Transit Authority (MTA)'s MetroCard system in New York City (Barry et al, 2002; Barry et al, 2009) revealed the feasibility of station-to-station OD matrix generation in the entry-only automatic fare collection subway system. Zhao et al. (2007) and Rahbee (2009) proposed a transit OD matrix estimation algorithm for origin-only AFC data from Chicago Transit Authority rail system. However, their algorithms primarily focused on the rail system, where boarding at fixed stations are easier to locate than bus transit systems. Pelletier et al. (Pelletier et al., 2010) undertook a thorough literature review on transit smart card data usage, and they concluded that properly processing SC data can enhance the strategic, tactical, and operational performances for transit agencies Trépanier et al. (2007; 2009) conducted several studies on AFC system in the National Capital Region of Canada, and developed algorithms to extract travel information from SC transaction data for transit performance measures. They evaluated various transit statistics, and demonstrated the feasibility of developing of a transit performance measure system using SC data. Munizaga and Palma (2012) developed a disaggregate multimodal approach to infer passengers' alighting stops using smart card data and GPS data in Santiago, Chile.

Most of the aforementioned studies are based on the entry-only AFC system, where boarding information is known in advance. In several existing AFC systems with missing boarding stops, researchers incorporated other data sources to jointly infer boarding locations, such as Automated Passenger Counter (APC) data, schedule data and GPS data. Farzin (2008) outlined a process to construct an automated transit OD matrix based on smart card and GPS data in Brazil. Nassir et al. (2011) integrated APC data, GPS data, transit schedule data with AFC data to estimate the stop-level passenger origin and destination. Zhang el al. (2007) matched each passenger's boarding time for origin inference by recording bus arriving time using on-bus surveys, but their algorithms are difficult to expand due to massive manual data collection effects. Review of existing literature does not identify any approach suitable for passenger OD information extraction from Beijing's SC transaction data. Hence, an OD estimation algorithm applicable for Beijing's AFC system is highly desired.

2.2 Travel Behaviors Analysis in Public Transit

Traditional travel survey or diary study is very costly and difficult to understand the transit rider's trip information and travel regularity. Transit passive data collection methods shed light on the individual-level transit behavior analysis on a multi-day basis. Recently, using smart card data to mine transit riders' travel patterns has been gaining more and more popularities, and a wealth of relevant researches have been conducted.

Pelletier et al. (2011) summarized previous smart card data studies, and showed that modeling individual based trip behavior is a potentially challenging topic. Kitamura et al. (13) and Morency et al. (2006;2007) utilized multiple day smart card data to analyze transit riders' travel variability, and concluded that understanding travel variability can reduce operational cost and manage demand. Several studies (2004;2005) concluded that transit agencies are able to extract the customer loyalty on a basis of multiday smart card data. Utsonomiya et al. (2006) used the smart card data from the Chicago Transit Authority (CTA) to extract passengers' transit usage and access distance, and concluded the transit usage data can be used for transit planning and market research. Webb (2010) emphasized the importance of transit loyalty, and developed several measures (e.g. satisfaction, quality of service) to quantify transit loyalty; however, Webb's findings are still based on traditional customer satisfaction survey. Lee and Hickman (2011) defined regular transit users as two or more trips during typical weekdays, and found travel patterns vary by card types. Lu and Reddy (2012) identified the irregularity of transit ridership by mining transit smart card data in New York City. Their findings will help transit agencies to optimize the weekday transit schedule for cost savings. Devillaine et al. (2012) took advantage of both smart card data and GPS data to extract transit riders' behavioral information such as activity location and time, duration, trip purpose.

Most of the aforementioned research based on smart card data extracted travel behavior information macroscopically rather than by analyzing individual transit riders' travel patterns. Chu and Chapleau (2010) applied the association rule and clustering algorithms to measure transit riders' regularity, and conducted an individual travel behavior analysis using both temporal and spatial methods. However, their analysis was based on high quality data with complete information and their method was not optimized for a large dataset. In reality, most transit agencies have adopted a comprehensive procedure to store smart card data, providing strict authorization and security mechanisms to protect the personal information generated from smart card data (Dinant and Keuleers, 2004). Sensitive content such as passenger age, name, boarding and alighting locations are intentionally truncated to address privacy concerns (Verykios et al., 2004), so efficient data mining approaches are needed to infer passenger travel behavior information from these incomplete smart card datasets.

2.3 Transit Performance Monitoring and Visualization

According to the Transit Capacity and Quality of Service Manual (TCQSM) in 1999 (Kittelson and Associates, 1999), six performance indicators are recommended to evaluate the public transit system: service frequency, hours of service, service coverage, passenger loading, reliability, and transit vs. automobile travel time. Lem et al. (1994) took into account intermodal performance measures, and proposed the most common indicators as: operating cost per revenue vehicle hour, operating cost per passenger boarding, farebox revenue per operating cost, number of boarding passengers per revenue vehicle mile, and passenger revenue vehicle hour. However, the transit performance measures proposed by TCQSM and Lem et al. rely on the manually collected data either by surveys or onboard questionnaires on a single day, and there is no further information provided to monitor the transit system performance in a long term due to data limitation (Trépanier et al., 2009).

In the past decades, more and more transit agencies have begun to adopt AFC systems for fare collection, and thus the availability of transit smart card data improves significantly. Transit cooperative research program (TCRP) identified the AFC system is a potential data collection technique for large-scale transit performance measures (Kittelson & Associates et al. 2003). At the same time, TCRP also indicated that integrating both AFC and AVL data holds substantial promise for improving transit planning and operations (Furth et al., 2006). To better utilized AFC data and AVL data for transit performance, a variety of relevant studies have been conducted. Bertini and EI-Geneidy (2003) processed the archived transit data from bus dispatch system (BDS) in Tri-County Metropolitan Transportation District of Oregon (TriMet). This system includes AVL based GPS, automated passenger counters (APC) and traffic signal priority (TSP), radio communication and computer-aided dispatching technologies. The stop-level data generated by this BSD system are stored periodically, and requires significant efforts to process and generate meaningful statistics for performance monitoring. Bertini and EI-Geneidy then resorted to visualization tools to further analyze and present these tremendous data, and proposed several valuable indicators to quantify the transit service variability. Gallucci and Allen (2011) used the transit performance measures program of Chicago's Regional Transportation Authority (RTA) as an example, and elaborated the challenges and summarized the lessons. They categorized the following five areas to create the transit performance indicators: service coverage, service efficiency and effectiveness, service delivery, service maintenance and capital investment, and service level solvency. Based on the experience of Chicago RTA, Gallucci and Allen further emphasized that quantifying both customer satisfaction and asset condition could be the possible direction for transit performance measures.

Without powerful visualization techniques, disseminating transit performance indicators generated from massive smart card data in an intelligible manner could be very difficult. GIS is often considered an effective tool to convey the spatial information, and a wealth of studies have been undertaken to incorporate passive transit data and GIS information to present transit performance measures or develop transit traveler information systems. For instance, Chapleau et al (2011) integrated AFC, APC, GPS and GIS data to establish a framework to evaluate transit performance in a GIS platform. Curries and Mesbah (2011) developed a GIS visualization platform to explore the spatial and temporal patterns of changes for transit performance in Melbourne, Australia. They utilized the ArcGIS software to demonstrate transit performance in a GIS environment. Liao and Liu (2010) developed a stand-alone visualization software interface to conduct the time point level travel time and schedule adherence analysis by using AVL, APC and AFC data in Minnesota. The processed information can be integrated into a mapping system to improve public transit service and optimize the transit route/schedule. The key performance components include: travel time analysis; reliability analysis for schedule adherence; travel time variation.

The majority of previous transit performance monitoring systems largely relies on the file-based data processing, and resort to commercialized GIS software to present the transit-related information for operations and planning. Transit agencies have to spend considerable time and financial resources purchasing and maintaining the software (Sun et al, 2011). In addition, because most of commercialize software are not designed on a basis of open architecture, transit agencies have to strictly provide the spatial data in accordance with the format of GIS files by the commercialized software. These obstacles incur inevitable inconveniences and lack flexibility for both users and developers. Moreover, file-based data management systems has their inherent disadvantages on processing tremendous amounts of data collected from AVL, APC and AFC systems, and disseminating the transit-related information in an efficient manner. To increase the interoperability and scalability, an open-architecture and open-source platform in consideration of geospatial database has been proposed and developed recently by Ma et al (2011). This new platform is named as Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net). DRIVE Net is not a simple system for data demonstration and archival, but also serves as a data-rich visualization platform, and it intends to take advantage of e-science developments for data-driven transportation research and applications. It is expected to remove the data barriers for better accessibility and extensibility, and benefit regular road users, transportation practitioners and researchers.

Chapter 3 Transit Passenger Origin Inference

Because smart card readers in the flat-rate buses do not record passengers' boarding stops, it is desired to infer individual boarding location using smart card transaction data. In this Chapter, two primary approaches are presented to achieve this goal. Approximately 50% transit vehicles are equipped with GPS devices in Beijing entry-only AFC system. Therefore, a data-fusion method with GPS data, smart card data and GIS data are firstly developed to estimate each bus's arrival time at each stop and infer individual passenger's boarding stops. And then, for those buses without GIS devices, a Bayesian decision tree algorithm is proposed to utilize smart card transaction time and apply Bayesian inference theory to depict the likelihood of each possible boarding stop. In order to expand the usability of proposed Bayesian decision tree algorithm in large-scale datasets, Markov chain optimization is used to reduce the algorithm's computational complexity. Both two transit passenger origin inference algorithms are validated using external data (e.g. on-board survey data and GPS data).

3.1 Passenger Origin Inference with GPS Data

In the first step, a GPS-based arrival information inference algorithm is presented to estimate the arrival time for each transit stop, and then, the inferred stop-level arrival time will be matched with the timestamp recorded in AFC system. The temporally closest smart card transaction record will be assigned with each known stop ID. The logic flow chart is demonstrated in Figure 3-1. The major data processing procedure will be detailed below.

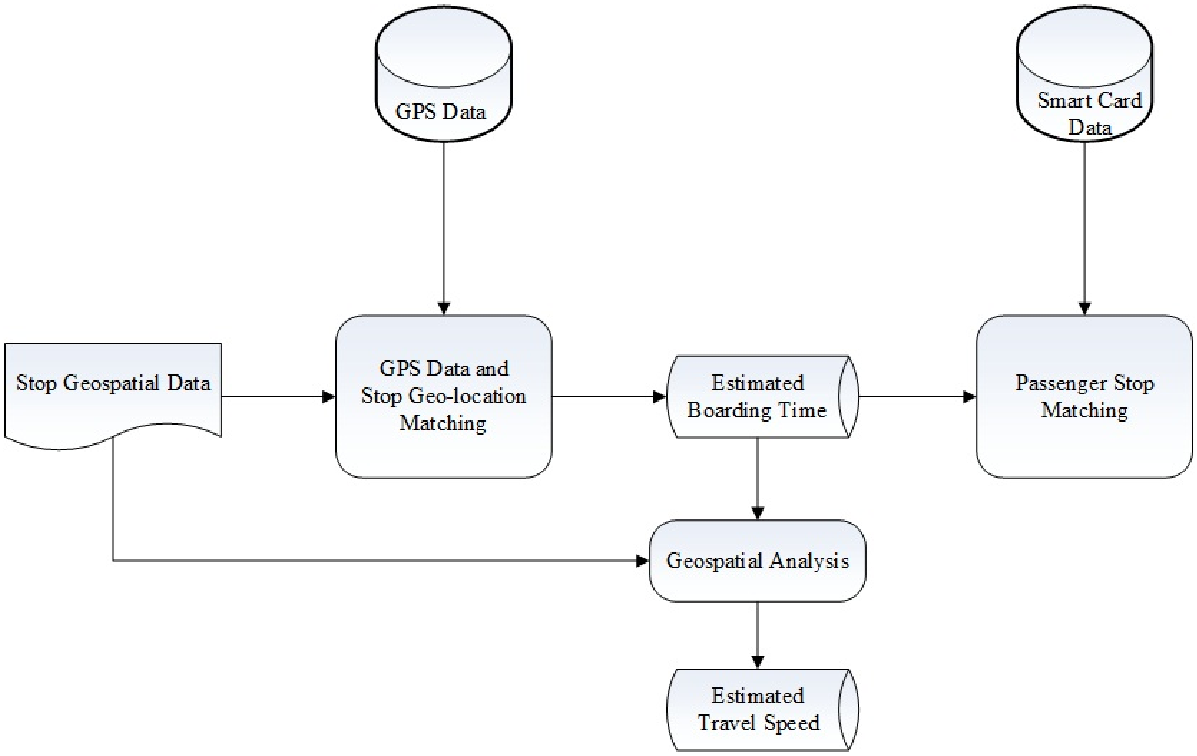


Figure 3-1 Flow Chart for Passenger Origin Inference with GPS Data

3.1.1 Bus Arrival Time Extraction

Three primary data sources are involved in the passenger information extraction: vehicle GPS data; transit stop spatial location data; and flat-fare-based smart card transaction data. A transit GIS network contains the geospatial location of each stop for any transit routes. The GPS device mounted in the bus can record each bus’s location and timestamp every 30 seconds, but the data quality of collected GPS records is insufficient: No directional information is recorded in Beijing AVL system; GPS points are off the roadway network due to the satellite signal fluctuation. Data preprocessing is required prior to bus arrival time estimation. A program is written to parse and import raw GPS data into a database in an automatic manner. Key fields of a GPS record are shown in Table 3-1.

Table 3-1 Examples of GPS raw data

Vehicle ID	Date time	Latitude	Longitude	Spot speed	Route ID
00034603	2010-04-07 09:28:57	39.73875	116.1355	9.07	00022
00034603	2010-04-07 09:29:27	39.73710	116.1358	14.26	00022
00034603	2010-04-07 09:29:58	39.73592	116.1357	19.63	00022
00034603	2010-04-07 09:30:28	39.73479	116.1357	0	00022
00034603	2010-04-07 09:30:58	39.73420	116.1357	3.52	00022

The first step is to estimate the bus arrival time for each stop by joining GPS data and the stop-level geo-location data. A buffer area can be created around each particular stop for a certain transit route using the ArcGIS software. Within this area, several GPS records are likely to be captured. However, identifying the geospatially closest GPS record to each particular stop is challenging since there could be a certain number of unknown directional GPS records within the specified buffer zone. Thanks to the powerful geospatial analysis function in GIS, each link (i.e. polyline) where each transit stop is located is composed of both start node and end node, and this implies that the directional information for each GPS record is able to infer by comparing the link direction and the direction changes from two consecutive GPS records. With the identified direction, the distance from each GPS point to this particular stop can be calculated, and the timestamp with the minimum distance will be regarded as the bus arrival time at the particular stop. Figure 3-2 visually demonstrates the above algorithm procedure. Inbound stop represents the physical location of a particular transit stop, and this stop is snapped to a transit link, whose direction is regulated by both a start node and an end node. By comparing the driving direction from GPS records with the link direction, the nearest GPS records to this particular stop can be identified, and marked by the red five-pointed star on the map. The timestamp associated with this five-pointed star will be considered as the estimated arrival time for this inbound stop.

The merit of the bus arrival time estimation algorithm lies in its efficiency. Rather than searching all the GPS data to identify the traveling direction for each stop, the proposed algorithm shrinks down the searching area, and filters out those unlikely GPS data. The operation greatly alleviates the computational burden, and is relatively easy to implement in the large-scale datasets, which is particularly critical to process the tremendous amount of datasets within an acceptable time period.

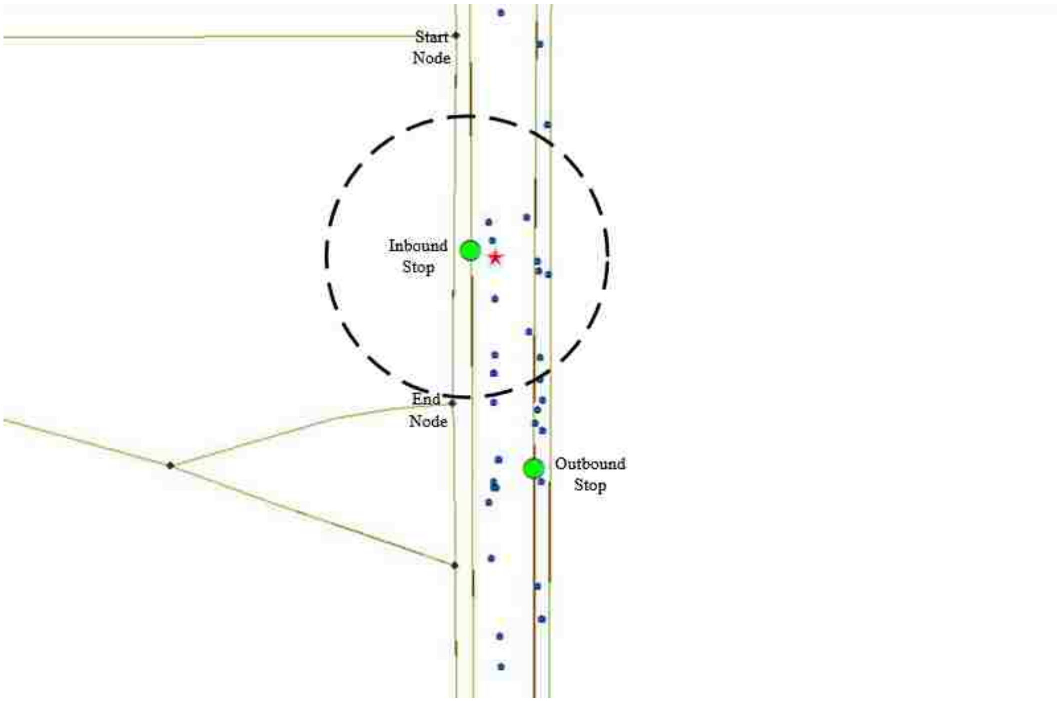
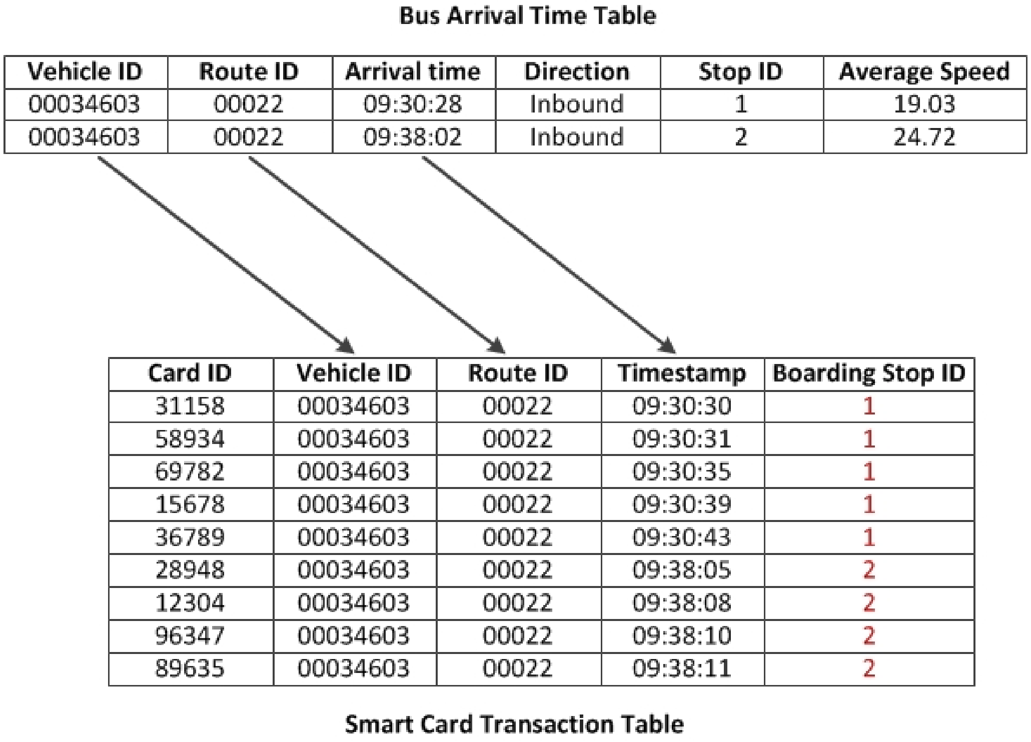


Figure 3-2 Boarding Time Estimation with GPS Data and Transit Stop Location Data

3.1.2 Passenger Boarding Location Identification with Smart Card Data

For each smart card data transaction record, the boarding stop can be estimated by matching the recorded timestamp and identified bus arrival time. As presented in Figure 3-3, for each smart card transaction record, the transaction time is compared with the inferred bus arrival time at each stop. This record will be assigned to a particular stop where the bus arrival time is the closest to the transaction time. Since passengers board the bus in a relatively short time interval, this data fusion method is able to capture almost all missing boarding stops.



Bus Arrival Time Table

Vehicle ID	Route ID	Arrival time	Direction	Stop ID	Average Speed
00034603	00022	09:30:28	Inbound	1	19.03
00034603	00022	09:38:02	Inbound	2	24.72

Card ID	Vehicle ID	Route ID	Timestamp	Boarding Stop ID
31158	00034603	00022	09:30:30	1
58934	00034603	00022	09:30:31	1
69782	00034603	00022	09:30:35	1
15678	00034603	00022	09:30:39	1
36789	00034603	00022	09:30:43	1
28948	00034603	00022	09:38:05	2
12304	00034603	00022	09:38:08	2
96347	00034603	00022	09:38:10	2
89635	00034603	00022	09:38:11	2

Smart Card Transaction Table

Figure 3-3 Boarding Stop Identification with Bus Arrival Time

In addition, because all the arrival time for all stops of a particular transit route can be estimated, the average travel time between two adjacent stops can be calculated as well and can be used to determine the average travel speed. Speed statistics are not only critical for transit performance measures, but also provide prior information for passenger origin inference when GPS data are absent.

3.1.3 Validation

Compared with bus arrival time, door opening time can be more accurately matched with smart card transaction time. This is because each bus may not exactly stop at each transit stop for passenger boarding. The inferred bus arrival time is subject to incur errors when it is used to match with smart card data. To validate the accuracy of the proposed data fusion algorithm for passenger origin inference, an on-board transit survey was undertaken to collect bus door opening time and arrival location for each stop of route 651 on January, 13th, 2013. Hand holding GPS devices were distributed to several volunteers to manually track the geospatial location of moving buses every 15 seconds. The survey duration was from 8:00 AM to 1: 00 PM, and a total of 75 bus door opening times were manually recorded. These bus door opening time records were then used to match the timestamps of smart card transactions from 417 passengers for boarding stop estimation, and these estimated stops can be considered as the ground-truth data. By comparing the ground-truth data with the results from the proposed GPS data fusion approach, 406 boarding stops were accurately inferred and 11 boarding stops differ from the ground-truth data within one-stop-error range. The proposed algorithm has an accuracy rate of 97.4%.

3.2 Passenger Origin Inference with Smart Card Data

There are still a fair amount of buses without GPS devices, and thus the bus arrival time at each transit stop is not directly measured. However, most passengers scan their cards immediately when boarding and almost all passengers should complete the check-in scan before arriving to the next stop. This indicates that the first passenger’s transaction time can be safely assumed as the group of passengers’ boarding time at the same stop. The challenge is then to identify the bus location at the moment of the SC transaction so that we can infer the onboard stop for that passenger. However, this is not easy because the SC system for the flat-rate bus does not record bus location. We know the time each transaction occurred on a bus of a particular route under the operation of a particular driver, but nothing else is known from the SC transaction database. Nonetheless, we are able to extract boarding volume changes with time and passengers who made transfers. By mining these data and combining transit route maps, we may be able to accomplish our goal. Therefore, a two-step approach is designed for passenger origin data extraction: smart card data clustering and transit stop recognition. To implement the proposed algorithm in an efficient manner, a Markov Chain based optimization approach is applied to reduce the computational complexity.

3.2.1 Smart Card Data Clustering

Transaction Data Classification

First of all, we need to sort SC transactions by the transit vehicle number. This results in a list of SC transactions in the vehicle for the entire period of operations for each day. During the operational period, the vehicle may have two to ten round-trip runs depending on the round-trip length and roadway condition. At a terminal station, a transit vehicle may take a break or continue running. So there is no obvious signal for the end of a trip (a trip is defined as the journey from one terminus to the other terminus). Meanwhile, there are a varying number of passengers at each stop, including some stops with no passengers.

For stops with several passengers boarding, all transactions can be classified into one group based on interval between their transactions. Thus, the clustered SC transactions can be represented by a time series of check-in passenger volumes at stops as shown in Table 3-2.

Table 3-2 Examples of Clustered SC transactions

Transaction Cluster No.	Stop ID	Stop Name	Total Transactions	Transaction Timestamp	Time Difference
1	Unknown	Unknown	18	5:26:36	0:14:26
2	Unknown	Unknown	9	5:41:02	0:03:16
3	Unknown	Unknown	11	5:44:18	0:04:35
4	Unknown	Unknown	27	5:48:53	0:01:00

In Table 3-2, total transactions indicate the total boarding passengers in one stop; transaction timestamp is recorded as the time when the first passenger boards in this stop, and time difference means the elapsed time between the boarding time at this stop and next stop with boarding passengers. Unlike most entry-only AFC systems in the United States, stop name and ID from each transaction are unknown in Beijing's AFC system. Most buses in service follow the predefined order of stops, however, it is still possible that there is no passenger boarding at a specific stop, and thus two consecutive SC transaction clusters do not necessarily correspond to two physically consecutive stops. Obviously, this further complicates the situation and the algorithm needed is indeed to map each cluster into the corresponding boarding stop ID.

In summary, the smart card data clustering algorithm contains three steps as follows:

1. All transaction data for each bus are sorted by the transaction timestamp in an ascending order.

2. For two consecutive records, if their transaction time difference is within 60 sec, then, these two transactions are included in one cluster; otherwise, another cluster is initiated.

3. If the transaction time difference for two consecutive records is greater than 30 min or driver changing occurs, it is likely that the bus has arrived in terminus, and for this bus, one bus trip has completed. Next record will be the beginning for the next bus trip.

The result of the clustering process is several sequences of clustered transactions. Each sequence may contain one or more trips of the transit vehicle. For particular routes, due to the limited space in terminus or busy transit schedule, bus layover time may be too short to be used as a separation symbol for trips. Such buses may have a very long clustered sequence that makes the pattern discovery process very challenging. Furthermore, unfamiliar passengers or passengers boarding from the check-out doors (this happens for very crowded buses) may take longer than 60 seconds to scan their cards. The delayed transaction may cause cluster assignment errors. Again, this adds extra challenge to the follow-up passenger origin extraction process.

Transaction Cluster Sequence Segmentation

Beijing has a huge transit network with nearly 1,000 routes. It is quite common to see passengers transfer between transit routes. Through transfer activity analysis, we can further segment the clustered transaction sequence into shorter series to reduce the uncertainty in passenger OD estimation (Jang, 2010). Two key principles used in the transfer stop identification are:

(1) We assume the alighting stop in the previous route is spatially and temporally the closest to the boarding stop for the next route. This is reasonable because most passengers choose the closest stop for transit transfer within a short period of time (Chu, 2008). Assume a passenger k makes a transfer from route i to route j within n minutes. If route i is a distance-based-rate bus line or a subway line, then we can identify the transfer station that is also the boarding stop of route j. Even if both routes are flat-rate bus routes, if the transferring location is unique, we can still use the transfer information to identify the transfer bus stop ID and name. In this study, the transfer time duration n is 30 minutes, and the maximum distance between two transfer stops is 300 meters.

(2) We assume that both the alighting time and the boarding time for each particular stop is similar. In this case, we can substitute a passenger's boarding stop with another passenger's alighting stop. Assume a passenger k makes a transfer from route i to route j. If route j is a subway line, where both its boarding location and time are available, then we can estimate the passenger k's alighting stop of route i, and this alighting stop can be also considered as the boarding stop for those passengers who get on the bus at the same time.

Walk distance between the two stops should be taken into account for inferring the time when the flat-rate bus arrives at the transfer stop. However, several possible boarding stops may exist due to the unknown direction in the flat-rate smart card transaction, and thus additional data mining techniques are needed to find the boarding stop with the maximum likelihood. These data mining techniques will be detailed in the next section.

Based on the identified transfer stops, we can further segment the transaction cluster sequence into shorter cluster series. Each series is bounded by either the termini or the identified bus stops. The segmented series of transaction clusters will be used as the input for the subsequent transit stop inference algorithm.

3.2.2 Data Mining for Transit Stop Recognition

Bayesian Decision Tree Inference

If we treat each segmented series of transaction cluster as an unknown pattern, this unknown pattern can be considered as a sample of the sequential stops on the bus route. If every stop has several passengers for boarding, this unknown pattern is identical to the known bus stop sequence. Also, since distance and speed limit between stops are known, travel time between stops is highly predictable if there is no traffic jam. In reality, however, there may have varying distribution of passengers boarding at any given stop and roadway congestion may cost unpredictable delays. Therefore, the unknown pattern recognition is a very challenging issue. Once the unknown pattern is recognized, the boarding stop for any passenger becomes clear.

Bayesian decision tree algorithm is one of the widely used data mining techniques for pattern recognition (Janssens et al., 2006). Each node in the Bayesian decision tree is connected through Bayesian conditional probability, and the entire tree is constructed directionally from the root node to the leaf nodes. Applying this technique to the current problem, we can represent the known starting stop as the root. if we denote the current boarding stop ID at time step k as S_k, and at time step k+1, the next boarding stop ID as S_{k+1}, according to Bayesian inference theory (Bayes and Price, 1763), S_{k+1} can be calculated as:

$$S_{k+1} = \arg\max_j (\Pr(S_{k+1} = j \mid S_1, S_2 ... S_k)) \tag{3-1}$$

where $\Pr(S_{k+1} \mid S_1, S_2 ... S_k)$ =conditional probability of the next boarding stop being S_{k+1}, given the previous boarding stop sequence $S_1, S_2 ... S_k$.

A Bayesian decision tree represents many possible known patterns. We need to compute the probability for each known pattern to match the unknown pattern. By further observation, we can find due to the nature of transit route, the probability of passengers boarding at S_{k+1} at time step k+1 is only related to whether the last boarding stop was S_k at time step k. That is because if the transaction time and corresponding bus location for SC transaction cluster k is known, the next SC transaction cluster k+1 only relies on how fast the bus travels during the time period between SC transaction clusters k and k+1. In this case, a SC transaction series can be recognized as a Markov chain process. Markov chain is a stochastic process with the property that the next state only relies on the current state.

Therefore, S_{k+1} can be rewritten as:

$$S_{k+1} = \arg\max_j(\Pr(S_{k+1} = j \mid S_1, S_2 ... S_k)) = \arg\max_j(\Pr(S_{k+1} = j \mid S_k = i)) \\ \textit{subject to } i < j \tag{3-2}$$

The single-step Markov transition probability is defined as $\Pr(S_{k+1} = j \mid S_k = i)$, also denoted as p_{ij}, with i, j being the stop IDs. Without losing generality, we assume the bus is moving outbound with an increasing trend of stop ID toward the destination. Then the transition probability matrix Π can be simplified as:

$$\Pi = \begin{pmatrix} p_{11} & p_{12} \cdots & p_{1n} \\ p_{21} & p_{22} \cdots & p_{2n} \\ \vdots & \vdots & \vdots \\ p_{(n-1)1} & p_{(n-1)2\cdots} & p_{(n-1)n} \\ p_{n1} & p_{n2\cdots} & p_{nn} \end{pmatrix} = \begin{pmatrix} 1 - \sum_{i=2}^{n} p_{1i} & p_{12} \cdots & p_{1n} \\ 0 & 1 - \sum_{i=2}^{n} p_{2i} \cdots & p_{2n} \\ \vdots & \vdots & \vdots \\ 0 & 0_{\cdots} & p_{(n-1)n} \\ 0 & 0_{\cdots} & 1 \end{pmatrix} \tag{3-3}$$

where n=the total number of stops for the bus route. This transition probability matrix plays a vital role in determining the potential stop ID for the next time step.

Transition Matrix Generation

To recognize the unknown pattern, it is critical to develop a measure to quantify p_{ij}, the possibility of next boarding stop being stop j conditioned on the previous boarding stop being i. The higher p_{ij} is, the more likely the next SC transaction cluster corresponds to boarding passengers at stop j. In other words, p_{ij} represents the probability for the next SC transaction cluster timestamp being the bus boarding time at stop j. That is to say, the boarding time in stop j for cluster k+1 can be predicted based on the travel distance from stop i to stop j and average bus speed. Then, the calculated time can be used as an indicator to compare with the real transaction timestamp for cluster k+1. From this point, the average speed between stops i and j will be a key variable. If the timestamp for cluster k is t_k, and that for cluster k+1 is t_{k+1}, then, the bus travel time from time step k to time step k+1 is $t_{k+1} - t_k$, and the stop distance between stop j and stop i is D_{ij}, then, the average bus travel speed V_{ij} can be expressed as:

$$V_{ij} = \frac{D_{ij}}{t_{k+1} - t_k} \tag{3-4}$$

Where V_{ij} is a random variable depending on the traffic condition at the moment. V_{ij} is considered to be normally distributed, and its probability density function can be adopted to quantifying p_{ij}.

In the speed normal distribution, the mean travel speed μ_{ij} and standard deviation σ_{ij} can be calculated from all buses with GPS devices in the same route. Under this circumstance, the boarding time for each stop can be inferred by matching GPS data and stop location information. Using the inferred boarding time difference and distance between stop i and stop j, we can calculate the mean travel speed μ_{ij} and standard deviation σ_{ij} as a priori information. It is noteworthy that the speed mean and standard deviation are not dependent on GPS data, but can be also obtained by other data sources such as distance-based-rate SC transaction data. A sensitivity analysis further demonstrates the algorithm's robustness even with different speed data sources.

Then, the transition probability can be reformulated as:

$$\begin{aligned} p_{ij} &= \Pr(S_{k+1} = j \mid S_k = i) \\ &= \int_{z_{ij}-\mathit{\Delta}}^{z_{ij}+\mathit{\Delta}} \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)dz = \frac{1}{\sqrt{2\pi}} \exp(-z_{ij}^2/2) \cdot 2\mathit{\Delta}, \end{aligned} \tag{3-5}$$

where $Z_{ij} = \frac{V_{ij} - \mu_{ij}}{\sigma_{ij}}$, which is the standardized travel speed between stop j and stop i, Δ is a small increase value for travel speed, and it will not impact the algorithm result, since this is a common term for each transition probability. In practice, to avoid the fast growth of Bayesian decision tree, the transition probability can be bounded by a minimum probability to eliminate those unlikely stops during calculation.

Each element in transition matrix can be quantified in the same way as shown in Equation (5). With the complete transition matrix, the unknown pattern of SC transaction series can be recognized as:

$$\begin{aligned}
&[S_{k+1}, S_k, S_{k-1}, ..., S_1] \\
&= \arg\max_{S_1 ... S_{k+1}} \Pr(S_{k+1}, S_k, S_{k-1}, ..., S_1) \\
&= \arg\max_{S_1 ... S_{k+1}} \left(\Pr(S_{k+1} \mid S_k, S_{k-1}, ..., S_1) \Pr(S_k, S_{k-1}, ..., S_1) \right) \\
&= \arg\max_{S_1 ... S_{k+1}} \left(\Pr(S_{k+1} \mid S_k) \Pr(S_k \mid S_{k-1}) \cdots \Pr(S_2 \mid S_1) \right) \\
&= \arg\max_{S_1 ... S_{k+1}} (\prod_{n=1}^{k} \Pr(S_{n+1} = j \mid S_n = i)) \\
&= \arg\max_{S_1 ... S_{k+1}} (\sqrt[k+1]{\prod_{n=1}^{k} \Pr(S_{n+1} = j \mid S_n = i)}) \\
&= \arg\max_{S_1 ... S_{k+1}} (\overline{P}(k+1))
\end{aligned} \tag{3-6}$$

Here, $\overline{P}(k+1) = \sqrt[k+1]{\prod_{n=1}^{k} \Pr(S_{n+1} = j \mid S_n = i)}$ denotes the geometric mean probability of passengers boarding stop sequence at time step k+1. It is also the probability for the identified stop sequence to match the unknown pattern.

3.2.3 Algorithm Implementation and Optimization

Implementation

As mentioned in the previous sections, due to the nature of transaction data, several issues need to be addressed in the process of Markov chain based Bayesian decision tree algorithm:

1. Direction identification:

Beijing transit AFC system doesn’t log the travel direction information for each route. We need to determine whether the bus is traveling inbound or outbound before algorithm execution. The solution is that we construct two Bayesian decision trees in each direction. Then the probability of the most likely stop sequence from each of trees will be compared and the one with the highest path probability wins.

2. Outlier removal

As mentioned in the Smart Card Data Clustering section, in some cases, the delayed transactions impact the accuracy of clustering algorithm, and these abnormal transactions are also labeled as outliers. The principal difficulty is that two inconsistent SC transactions by timestamp that should be classified in one cluster may be read separately, and thus, the latter will be classified as another cluster for the next stop. For instance, at a particular stop, if one passenger boarded the bus and paid the fare at 8:00 AM, another passenger swiped his smart card to alight at 8:10 AM. Due to the relative large transaction timestamp gap, the second transaction will be assigned to another cluster. In this case, the boarding stop ID will be misidentified.

The strategy used to remove these outliers is that there exists a probability that a passenger may retain in the same stop. If the previous stop ID is defined as i, the number of total stops in each possible direction is denoted as N, and the probability that a passenger stay at stop i in the next time step can be expressed as:

$$p_{ii} = 1 - \sum_{j=i+1}^{j \leq N} p_{ij} \tag{3-7}$$

The probability is able to better depict the situation where passengers may delay a certain period to swipe their smart cards during boarding.

3. Bus trip detection

The journey begins from the initial bus stop to the terminus is defined as a bus trip. The bus terminus is designed for bus turning, layover, and driver change. It is also the starting stop on the bus timetable. However, in Beijing’s transit network, some bus termini are located in the busy street or have limited space. Hence, buses using these termini have to begin their next trip in a short time period without causing an obstruction. This is a challenging issue in the procedure of passenger origin inference, since the initial stop (root node) in Bayesian decision tree may be misidentified if the bus trip is mistakenly detected. The solution to this issue is to model the travel time probability of each transaction cluster series. As indicated in the transaction cluster sequence segmentation section, a transaction cluster sequence can be segmented by several series using aforementioned spatiotemporal transfer relationships. Each identified series is bounded by possible inferred stops, by calculating the travel time for multiple combinations of inferred stops, and comparing with the actual time difference, we are able to determine the existence of a bus trip based on the highest probability. Figure 3-4 demonstrates the procedure of identifying a bus trip.

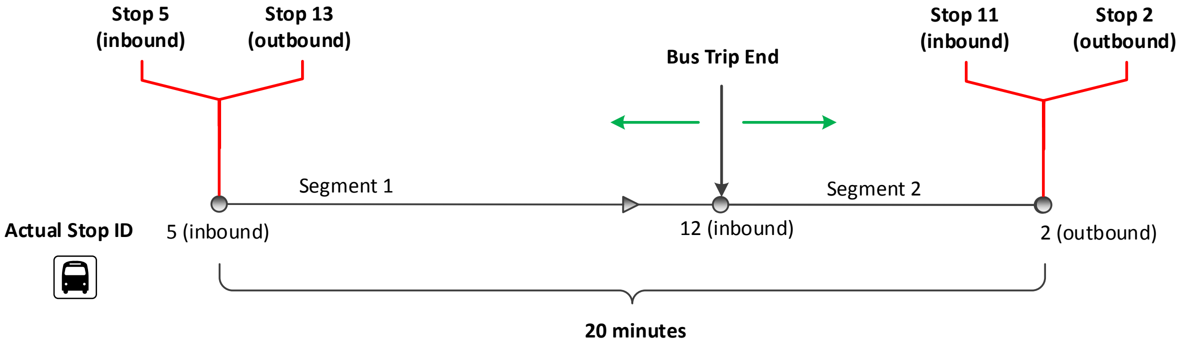


Figure 3-4 Bus Trip Identification

As presented in Figure 4-3, the starting point and ending point of a transaction sequence can be identified by several possible stops in different directions, and the duration of this transaction clustered sequence is known as 20 minutes. A variety of trips may exist for this transaction cluster sequence:

Trip 1: The bus travels from the 5th inbound stop to the 11th inbound stop.

Trip 2: The bus travels from the 5th inbound stop to the 2nd outbound stop.

Trip 3: The bus travels from the 13th outbound stop to the 11th inbound stop.

Trip 4: The bus travels from the 13th outbound stop to the 2nd outbound stop.

The maximum and minimum travel time for any trip can be obtained through GPS data or distance-based buses. In addition, the maximum bus layover time can be assumed as 30 minutes. According to the central limit theorem, the bus travel time in a known road segment should follow normal distribution, and therefore, we can compute the probability of each scenario, and choose the trip with the maximum probability. If the travel time from stop i to stop j is denoted as t_{ij}, and the probability density function of t_{ij} is defined as:

$$p(t_{ij}) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp(-\frac{(t_{ij} - \mu_{ij})^2}{2\sigma_{ij}^2})dt_{ij} \tag{3-8}$$

Where μ_{ij} is the average travel time from stop i to stop j, and σ_{ij} is the standard deviation of travel time from stop i to stop j. If the maximum and minimum travel time (plus maximum and minimum bus layover time) between stop i to stop j are $\max(t_{ij})$ and $\min(t_{ij})$ respectively, then the 95% confidence interval of travel time can be further expressed as:

$$[\mu_{ij} - 1.96\sigma_{ij}, \mu_{ij} + 1.96\sigma_{ij}] = [\min(t_{ij}), \max(t_{ij})] \tag{3-9}$$

The probability density function of t_{ij} can be rewritten as:

$$p(t_{ij}) = \frac{1}{\sqrt{2\pi(\frac{\max(t_{ij}) - \min(t_{ij})}{3.92})^2}} \exp(-\frac{(t_{ij} - \frac{\max(t_{ij}) + \min(t_{ij})}{2})^2}{2(\frac{\max(t_{ij}) - \min(t_{ij})}{3.92})^2})dt_{ij} \tag{3-10}$$

Each probability for the above four trips can be calculated as 0.54, 0.87, 0.0003,0 respectively. Therefore, the transaction cluster sequence starts at the 5th inbound stop, and ends at the 2nd outbound stop, and thus a terminus (bus trip end) should exist during this trip. This result matched with the actual bus trip. In these cases, the proposed Bayesian decision tree algorithm is able to identify each bus trip from the clustered transaction sequence.

Computational Performance Optimization

Although we illustrated the mathematical form for Markov chain based Bayesian decision tree in theory, this algorithm presented above has not been applied in the real dataset. Cooper (1990) has proven Bayesian decision tree algorithm a NP (Non-deterministic Polynomial)-hard problem, which means that this algorithm cannot be solved in a polynomial time. Conventional approach to calculate the path probability for all the potential boarding stop sequences is computationally expensive, especially for the long sequences. To better explain this challenge, an example is shown as follows:

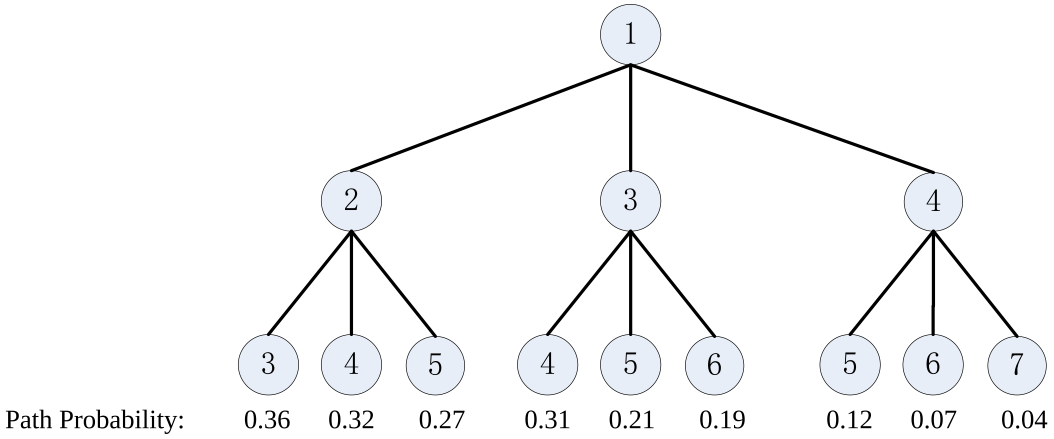


Figure 3-5 A Bayesian Decision Tree Algorithm Example

Assume the initial boarding stop is 1. The potential stops in the next step could be stop 2, stop 3, or stop 4 because they are all in the reachable range. Assuming that the situations are similar for the remaining stops, a decision tree is fully established. The traditional exhaustive search is to traverse each potential path, and select the maximum probability. Based on this method, we need to calculate the path probability nine times. This implies that the number of paths to be calculated increases exponentially as the time step increases. However, at the time step 3, there are two or more paths ending with stop 3, 4 and 5. Before carrying on the computation in the next time step, we can compare the probability of the paths with the same ending stop, and choose the maximum one, which is also called the partial best path, that is:

In the time step 3, only the following five paths are selected 1->2->3, 1->2->4,1->2->5,1->3->6, and 1->4->7. Recall that the Markov Chain model states that the probability of current state given a previous state sequence depends only on the previous state. Hence, five paths calculated in time step 3 guarantees the most probable paths in time step 4 without extra computations of other paths. According to Equation (3-11), we can express the optimized procedure in mathematics as:

$$\overline{P}(k+1) = \max_{i,j}(\overline{P}(k)(\sqrt[k+1]{\Pr(S_{k+1} = j \mid S_k = i)})) \tag{3-11}$$

We can now calculate the probability at each time step recursively until the end of the route. Computing the probability in this way is far less computational expensive than calculating the probabilities for all sequences. If we denoted the total stops for a specific route as n, and the SC transactions are classified in m clusters, which correspond to m time steps in Bayesian decision trees, then the computational complexity for the exhaustive approach can be written as $O(m^n)$. While using the optimized algorithm, the computational complexity is only $O(mn)$. With the optimization, the algorithm can be solved in a finite time, and can be efficiently applied in reality.

3.2.4 Validation

By installing GPS receivers on flat-rate buses, we can collect the geospatial information and spot speed data in a real-time manner. There are approximately 50% buses equipped with GPS devices in Beijing, and GPS data are updated every 30 seconds. These data provide the opportunity to validate the Markov-chain based Bayesian decision tree algorithm developed in this study for passenger origin data extraction. GPS coordinates and timestamp can be used to determine bus boarding and alighting location and time respectively. First, the geographical feature of bus stops and consecutive GPS records for each bus are joined using latitude and longitude coordinates. Then, by matching the passenger check-in time in the SC transaction database, the boarding stop ID can be associated with each transaction. Since the inferred stop ID using GPS data have been validated using the bus on-board survey method, and can be considered as the ‘ground truth’ data for the comparison purpose.

In this section, the Markov chain based Bayesian decision tree algorithm is first validated using GPS data for route 22, and then, several sensitivity analyses are conducted to investigate impacts of different parameter settings in Bayesian decision tree. Finally, a computational complexity experiment is also included at the end of this section.

Algorithm Validation

Flat-rate based route 22 was selected to infer unknown boarding location using Markov-chain-based Bayesian decision tree algorithm, and GPS data associated with route 22 were also collected to verify the result. The SC transaction data and GPS data were all recorded on April 7, 2010. The minimum stop probability is defined as 0.05. If a stop whose transition probability is less than 0.05, then this stop will be abandoned. Route 22 contains a total of 34 inbound and outbound stops as shown in Figure 3-6.

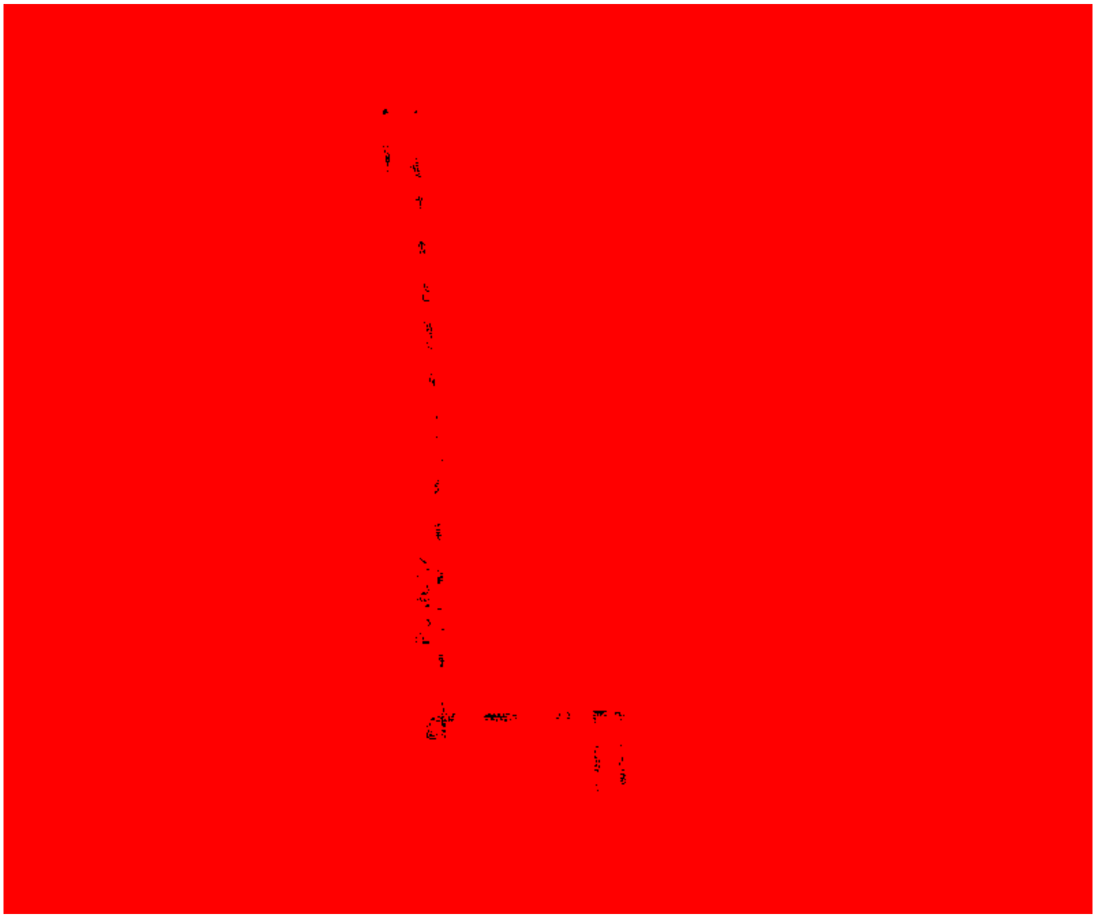


Figure 3-6 Route 22 in Beijing Transit Network

The algorithm results are listed as in Table 3-3 and Figure 3-7. In Table 3-3, there are a total of 12,675 SC transactions mapped with GPS data for Route 22. Error is defined as the stop ID difference (two stops that are adjacent to each other should have consecutive IDs) between the ground truth stop based on GPS data and the inferred stop using the proposed algorithm. For Route 22, 95% passenger boarding stops were deducted by the proposed algorithm. 55.8% of results perfectly matched with the stops inferred by GPS accurately. There are 11,645 recognized boarding stops within three-stop distance away from the actual boarding stop, accounting for approximately 96.7% of the total identified records or 91.6% of total records.

Table 3-3 Results of Bayesian Decision Tree Algorithm for Route 22 Based on GPS Speed

Route 22	Number of records	Accumulated percentage in inferred records	Accumulated percentage in total records
Stop ID error<1	7062	58.6%	55.8%
Stop ID error<2	10371	86.1%	81.8%
Stop ID error<3	11341	94.2%	89.5%
Stop ID error<4	11645	96.7%	91.9%
Total	12043	N/A	97.9%

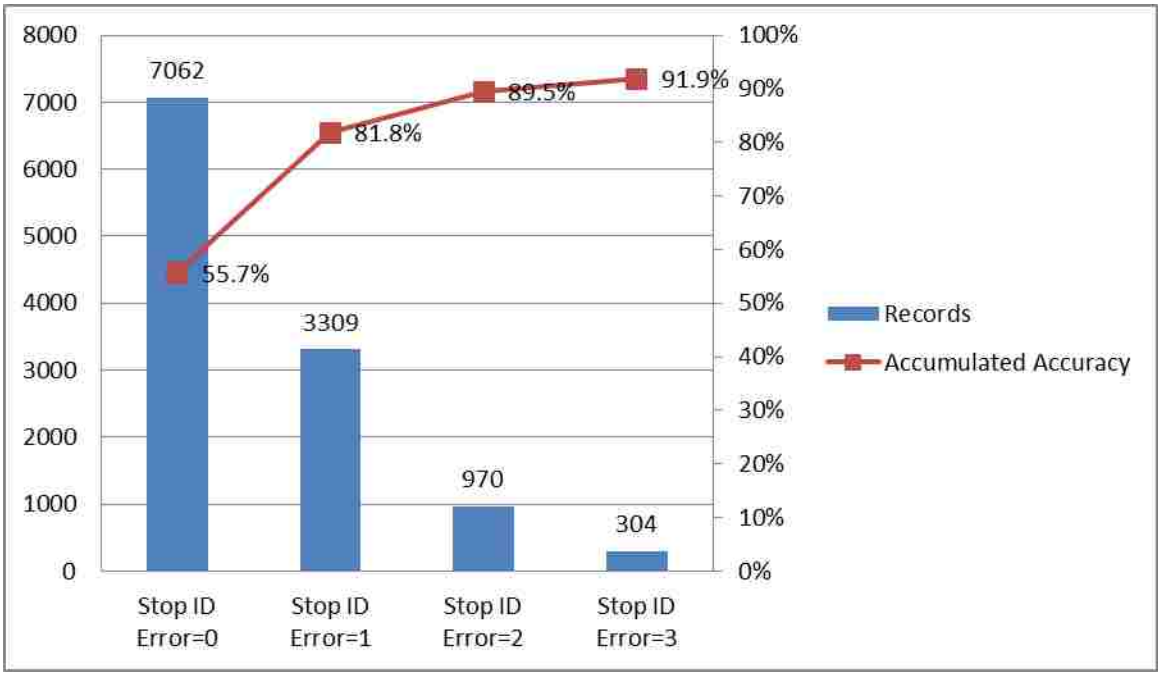


Figure 3-7 Bayesian Decision Tree Algorithm Accuracy for Route 22 Based on GPS Speed

The results are very encouraging. In Beijing’s transit network, the error within three stops is acceptable for transit planning level study, since these stops are mostly affiliated with the same traffic analysis zone (TAZ) due to the high transit network density.

Sensitivity Analysis

(1) source of travel speed calculation

Recall that in computing the transition matrix, mean travel speed μ and standard deviation σ were extracted from GPS data. However, there are still many flat-rate routes without GPS devices. To understand how the algorithm result changes when the travel speed mean and standard deviation are inaccurate, a sensitivity analysis is carried out for this purpose. Table 3-4 and Figure 3-8 show the results when the mean and standard deviation of travel speed are retrieved from the distance-based fare routes, and these routes share common stops with the "no-GPS" flat-fare route. Because both boarding stop and alighting stop are known in the distance-based fare buses, we are still able to extract the mean and standard deviation of travel speed between adjacent stops for transition matrix construction.

Table 3-4 Results of Bayesian Decision Tree Algorithm for Route 22 Based on Speed from Distance-based Fare Routes

Route 22	**Number of records**	**Accumulated percentage in inferred records**	**Accumulated percentage in total records**
Stop ID error<1	6841	58.5%	54%
Stop ID error<2	10319	88.2%	81.4%
Stop ID error<3	11296	96.6%	89.1%

Stop ID error<4	11509	98.4%	90.8%
Total	11694	N/A	92.2%

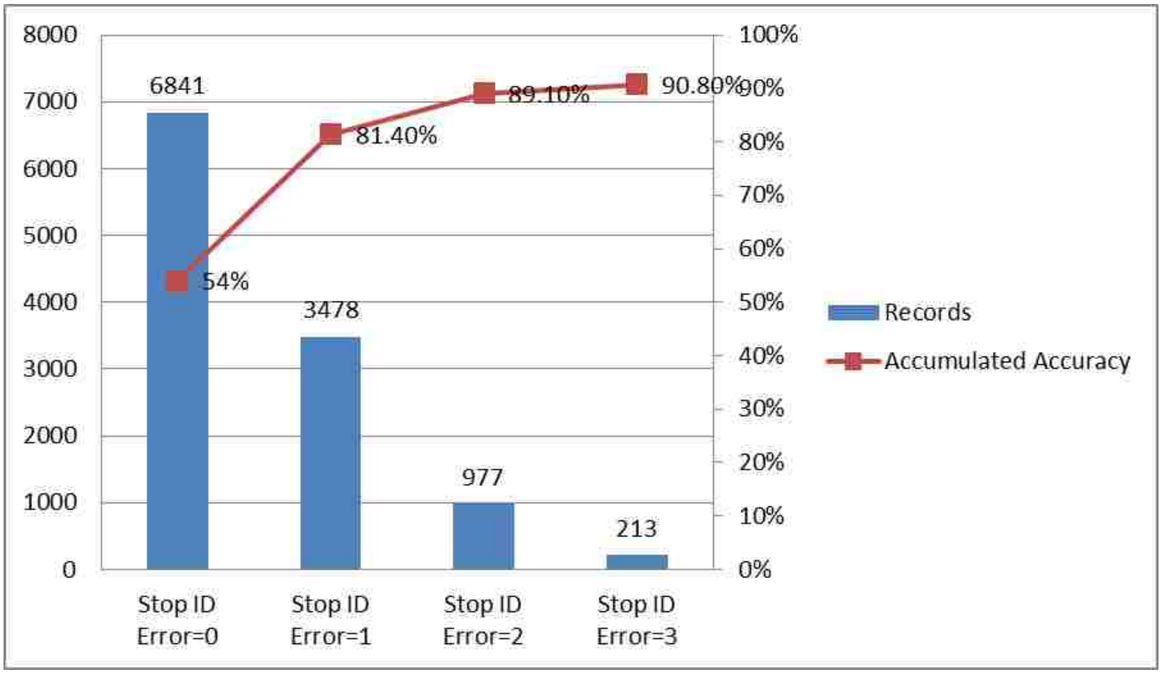


Figure 3-8 Bayesian Decision Tree Algorithm Accuracy for Route 22 Based on Speed from Distance-based Fare Routes

Different data sources only slightly influence the percentage of inferred stops. 92.2% boarding stops can be estimated using the speed generated from distance-based fare routes, and the accuracy within three-stop error is 90.8%. The result indicated the proposed algorithm is not sensitive to the travel speed, even without GPS data, we are still able to correctly identify passenger boarding stops using other data sources. This is not surprising, because in normal distribution, mean and standard only influence the shape for probability density function, as long as we make a reasonable assumption for bus travel speed calculation, the algorithm results will not fluctuate significantly.

(2) minimum stop probability

Minimum stop probability plays a vital role in impacting both the accuracy and efficiency of the proposed algorithm. A too high threshold may eliminate possible boarding stop candidates, and a too low threshold may consume additional computational resources. In this sensitivity analysis, a different minimum stop probability is set as 0.1, which means if the calculated transition probability of a particular stop is lower than 0.1, and then this stop is considered as an unlikely boarding stop. The comparison result is presented in Table 3-5 and Figure 3-9:

Table 3-5 Results of Bayesian Decision Tree Algorithm for Route 22 with Minimum Stop Probability as 0.1

Route 22	Number of Records	Accumulated Percentage in inferred records	Accumulated Percentage in total records
Stop ID error<1	6011	55.2%	47.4%
Stop ID error<2	9157	84.0%	72.2%
Stop ID error<3	10139	93.1%	80.0%
Stop ID error<4	10589	97.2%	83.5%
Total	10894	N/A	85.9%

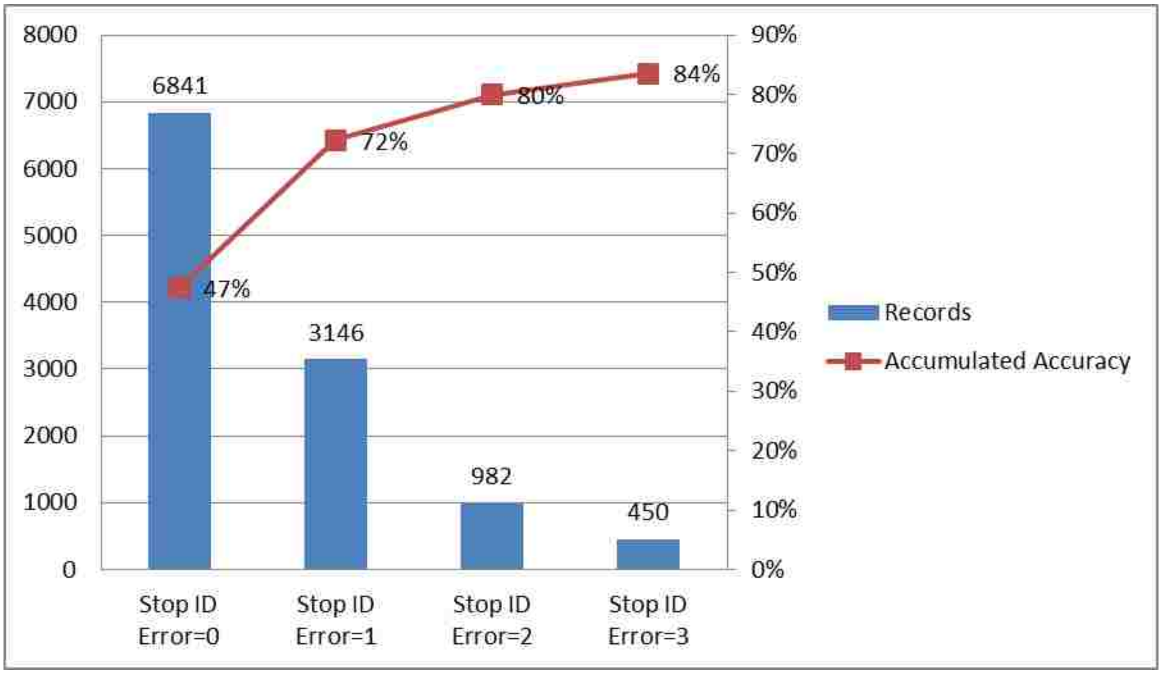


Figure 3-9 Bayesian Decision Tree Algorithm Accuracy for Route 22 with Minimum Stop Probability as 0.1

When the minimum stop probability increases, less boarding stops can be inferred using the proposed algorithm. In addition, the inferred boarding stops are less accurate compared with the ones with minimum stop probability as 0.05. This is a reasonable result since a rigorous probability threshold may limit the prorogation of errors. However, a trade-off exists between algorithm accuracy and efficiency.

Computational complexity comparison

As mentioned in the algorithm optimization section, the computational complexity should be also taken into account when the proposed algorithm is implemented in a large-scale transit network. To compare the algorithm efficiency between the basic Bayesian decision tree algorithm (Basic BDC) and the Markov chain based Bayesian decision tree algorithm (Markov-chain BDC), seven transit routes with an increasing number of total stops are tested. 10,000 smart card transactions for each route on April, 7, 2010 are used for comparison purposes. The experimental result is listed in table 3-6 and figure 3-10.

Table 3-6 Computational Complexity Comparison between Basic and Markov-chain Based Bayesian Decision Tree Algorithms

Route ID	Number of stops	Run time for Markov-chain BDC (milliseconds)	Run time for Basic BDC(milliseconds)
00616	23	3798	493740
00647	36	4890	674820
00005	53	7747	937387
00839	66	17082	1947348
00355	74	21071	2486378
00646	80	23979	4556010
00603	86	29114	5560774

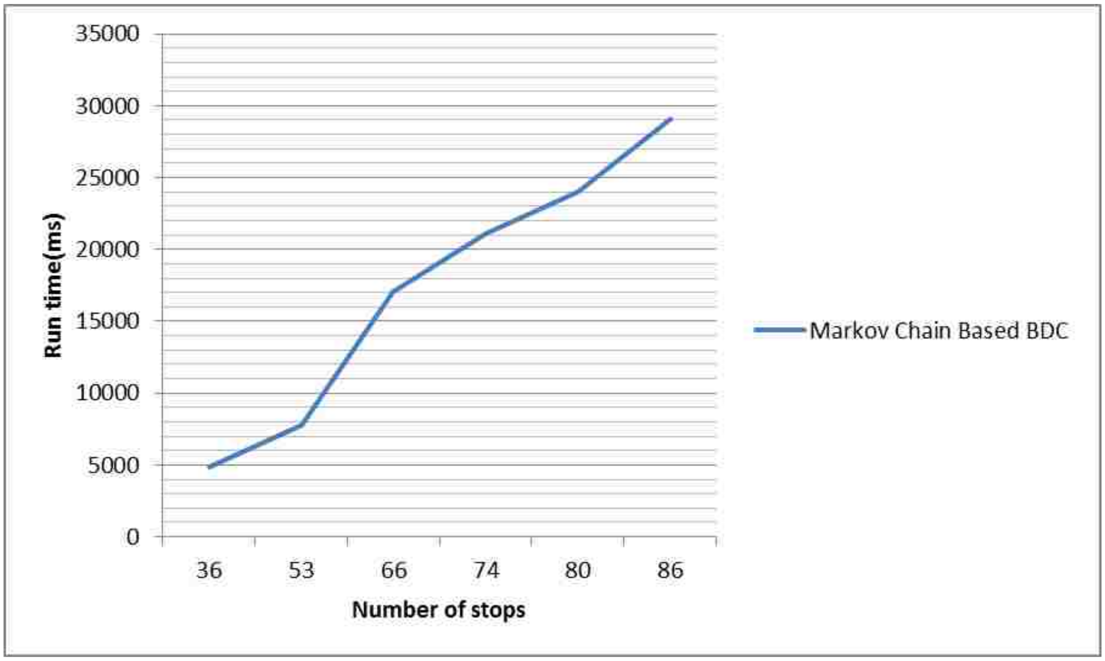


Figure 3-10 Markov Chain based Bayesian Decision Tree Algorithm Run Time Analysis

The Markov chain based BDC algorithm can save a significant amount of run time compared with the Basic BDC algorithm. The average performance gains can achieve to 142 times faster than the basic algorithm. This is because most of the redundant calculation steps have been already excluded using Markov chain property.

3.3 Conclusion

Different from most entry-only AFC systems in other countries, Beijing's AFC system does not record boarding location information when passengers embark the buses and swipe their smart cards. This creates challenges for passenger OD estimation.

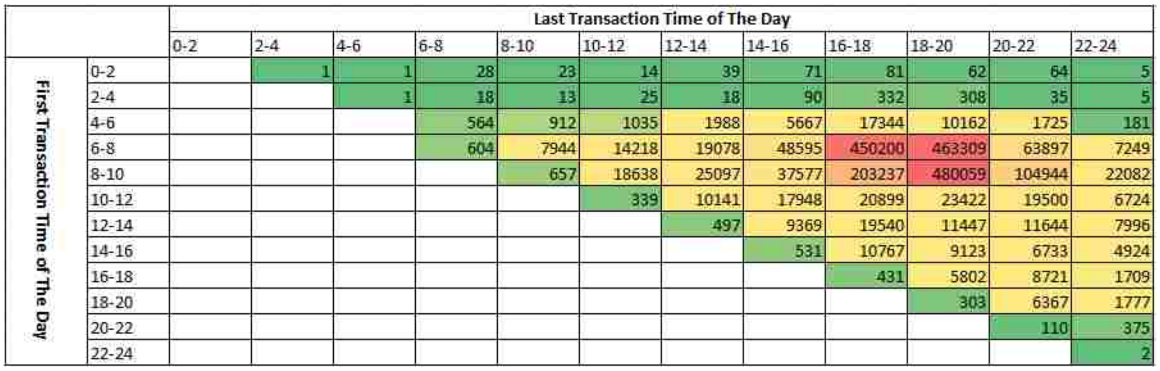
This chapter aims to tackle this issue. With further investigations on SC transactions data, we proposed a Markov chain based Bayesian decision tree algorithm to infer passengers boarding stops. This algorithm is based on Bayesian inference theory, and the normal distribution of travel speed between adjacent stops is used to depict the randomness of passenger boarding stops. Both the mean and the standard deviation can be obtained from GPS data or distance-based fare routes. Moreover, stationary Markov chain property is also incorporated to further reduce the computational complexity of the proposed algorithm to linear. The optimized algorithm is proven its accuracy using the SC transaction data.

This algorithm can be improved in various ways; for instance, the algorithm does not perform well under the circumstance that the travel speed between adjacent stops is not distinct, i.e. the travel speed probability calculated for each stop is similar. The potential countermeasure for this issue is to incorporate heterogeneity, e.g., the accessibility of a subway station or a central business district (CBD) for each transit stop.

In summary, the Markov chain based Bayesian decision tree algorithm provides both effective and efficient data mining approach for passenger origin data extraction. It sets up a great foundation to mine transit passenger ODs from the SC transaction data for transit system planning and operations.

Chapter 4 Transit Passenger Travel Pattern Mining

To demonstrate the temporal travel patterns and the pattern regularity for transit riders in Beijing, consider a typical travel week (in this case, the week of Monday July 5th to Friday July 9th, 2010). The transaction data from 3,845,444 smart cards was collected for that week, 58% of which (2,225,298 cards) contained two transactions for all five weekdays. FIGURE 1 shows the temporal frequency distribution of the "transaction pair" of the first transaction time and the last transaction time of the smart cards with two transactions per day. As shown in the red cells of FIGURE 1, most of the transit riders began their first trip between 6 AM and 10 AM, and ended their travel for the day between 4:00 PM and 8:00 PM. This is likely to represent a typical commuting trip chain, where a transit rider takes a bus or subway from his or her home to their place of work in the morning and then returns home in the evening. The temporal distribution shown in the figure implies that strong temporal travel patterns exist in the multiday smart card data. However, the regular spatial travel pattern for a specific card holder remains uncertain and will be explored in the analysis described below.



		Last Transaction Time of The Day											
		0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24
First Transaction Time of The Day	0-2		1	1	28	23	14	39	71	81	62	64	5
	2-4			1	18	13	25	18	90	332	308	35	5
	4-6				564	912	1035	1988	5667	17344	10162	1725	181
	6-8				604	7944	14218	19078	48595	450200	463309	63897	7249
	8-10					657	18638	25097	37577	203237	480059	104944	22082
	10-12						339	10141	17948	20899	23422	19500	6724
	12-14							497	9369	19540	11447	11644	7996
	14-16								531	10767	9123	6733	4924
	16-18									431	5802	8721	1709
	18-20										303	6367	1777
	20-22											110	375
	22-24												2

Figure 4-1 Weekday Temporal Distribution for Transit Smart Card Holders with Two Transactions for the week of 5th-9th July 2010

The focus of this study is twofold: individual travel pattern recognition and travel regularity mining. A flow chart of the work performed for the study is illustrated in FIGURE 2: (1) retrieve each passenger’s multi-day’s smart card transactions from the database; (2) generate this passenger’s trip chains utilizing their spatiotemporal relationships; (3) apply a series of data mining approaches to extract this passenger’s travel pattern and travel regularity based on the generated trip chains. To reduce the complexity involved in the regularity clustering algorithm, association rules were identified for the large-scale smart card data mining process.

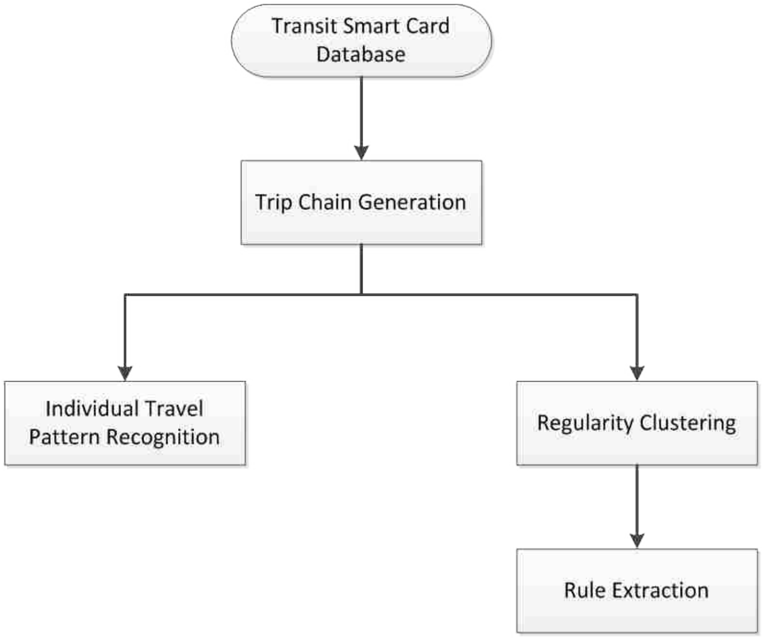

Transit Smart Card Database
Trip Chain Generation
Individual Travel Pattern Recognition
Regularity Clustering
Rule Extraction

Figure 4-2 Flow Chart of the Study Research Process

4.1 Trip Chain Generation

Before the spatial and temporal patterns of individual transit riders can be examined, their trip chain information must be constructed. A trip chain is defined as a series of trips made by a traveler on a daily basis and is considered a useful way to demonstrate travelers' behaviors (McGuckin and Nakamoto, 2004). For flat fare buses, transit riders swipe their smart cards only during boarding and the smart card reader is not able to record either their boarding location or when and where they alight. In order to estimate transit riders' boarding stops, a Markov chain based Bayesian decision tree algorithm by (Ma et al., 2012) was therefore utilized to extract changes in the boarding volume with time between two consecutive transactions and apply this information, in conjunction with historical speed profiles retrieved from GPS data, to calculate the probabilities for all potential boarding stops; the stop with the maximum probability was assumed to be the boarding stop. Based on this algorithm, more than 90% of the smart card data can be accurately assigned. For distance-based fare buses, although boarding times are not recorded by the smart card reader, other information such as each passenger’s boarding stop location, alighting stop location, and alighting time are known. Therefore, the missing boarding time can be approximately substituted by another passenger’s alighting time at the same stop.

A fixed temporal threshold was used in our study to link several smart card transaction records into a trip chain. Fixed temporal thresholds change depending on the type of transfer activity. For instance, if a passenger transfers from a distance-based fare bus to a flat fare bus, the alighting time in the previous trip and the boarding time in the current trip are known and the appropriate 30 minute time interval recorded. However, if this passenger makes a transfer from a flat fare bus to a distance-based fare bus, the alighting time is not recorded when he/she exits from the flat fare bus so a 60 minutes time interval was utilized to differentiate various trips in this study to take into account both in-vehicle travel time and transfer time. The determination of transfer time intervals for different transfer activities was based on the 2010 Beijing 4th Comprehensive Transport Survey (Beijing Transportation Research Center, 2012), with the average transit transfer time and in-vehicle travel time being 25.4 minutes and 40 minutes, respectively. The same survey revealed that more than 94% of the transfer activities took less than 60 minutes, so if the transaction time difference between two consecutive smart card records was greater than 60 minutes, a new trip was generated; times less than this were taken to represent a transfer activity between two routes or two transportation modes (bus and subway) (Jang, 2010).

Table 4-1 shows linked trip chain examples extracted from the study data. Here, *Chain ID* is a unique identifier for each trip chain sorted in ascending order by the transaction time. For each *Card ID*, the first trip's boarding time (*First Boarding Time)* and the last trip's alighting time (*Last Alighting Time*) are associated with that *Chain ID*. *Route Sequence* refers to the routes the rider took and *Stop ID Sequence* refers to the boarding and alighting stop IDs for distance-based fare buses. As previously noted, only distance-based fare buses and subways record both boarding and alighting locations, but the subway AFC system also has no check-out smart card scan reader when transit riders transfer between different lines. Take Chain ID 46388399 as an example. The transit rider boarded the distance-based fare bus on Route 635 at Stop ID 99964, and alighted at Stop ID 99966.That individual then made a transfer to subway Line 5 at Stop ID 50258, finishing his or her journey by exiting subway Line 10 at Stop ID 50167. Due to the lack of alighting location information for flat fare buses, some of the trip chains suffer from missing alighting time and stop ID sequence information, e.g. Chain ID 46388408 in TABLE 1. However, this does not have a huge impact on the accuracy of the individual travel pattern recognition and regularity clustering algorithms since both algorithms are capable of handling both missing values and outliers.

Table 4-1 Extracted Trip Chain Information for an Individual Transit Rider for the Week of 5th-9th July, 2010

Chain ID	Card ID	Date	First Boarding Time	Last Alighting Time	Route Sequence	Stop ID Sequence
46388399	1000751018309337	20100705	07:08:45	07:47:28	00635->10->13	99964,99966->50258,50167
46388400	1000751018309337	20100705	18:15:24	18:53:10	13->10->00635	50192,50245>100013,100015
46388401	1000751018309337	20100706	07:19:21	08:01:13	00350->10->13	91267,91269->50258,50167
46388402	1000751018309337	20100706	17:56:08	18:49:50	13->10->00635	50192,50245>100013,100015
46388403	1000751018309337	20100707	07:10:43	07:49:21	00635->10->13	99964,99966->50258,50167
46388404	1000751018309337	20100707	18:29:00	19:06:47	13->10->00350	50192,50245->91276,91278
46388405	1000751018309337	20100708	21:13:58	21:40:10	5->10	50125,50246
46388406	1000751018309337	20100709	07:16:24	08:03:46	00635->10->13	99964,99966->50258,50167
46388407	1000751018309337	20100709	17:25:00	18:11:59	13->10->00635	50192,50245>100013,100015
46388408	1000751018309337	20100709	18:30:31	NULL	00031	NULL

Note: Subway routes are denoted as one or two digits.

4.2 Individual Travel Pattern Recognition

Once the trip chain info has been constructed, the travel pattern for each transit rider is further investigated through clustering the trip chains. As shown by the example in TABLE 1, an individual transit rider is likely to show a certain travel pattern during a multi-day period. To retrieve these hidden and repeated travel patterns in an efficient manner, the density-based spatial clustering of application with noise (DBSCAN) algorithm was therefore adopted. Unlike most non-hierarchical clustering algorithms, the DBSCAN algorithm is not required to define the number of clusters (Ester et al., 1996) or identify arbitrarily shaped clusters because higher-density records are more likely to be grouped into a cluster. Two key parameters do, however, need to be defined in the DBSCAN algorithm: the ε distance and the minimum number of points (*MinPts*). The ε distance defines the density-reachable range; if a sample record falls within the ε distance, then this record will be included into an existing cluster. *MinPts* limits the minimum number of records in each cluster; if the number of records in each final cluster is less than *MinPts*, then these records are marked as noise. If the records are close to each other (i.e. more dense), these records are more likely to be clustered by DBSCAN. An outlier is often distinct from other dense records, so DBSCAN is able to detect these outliers.

A transit rider may begin their repeated trips in both the spatial and temporal domains and transit riders' recurring boarding/alighting locations and times are considered simultaneously for clustering. In our application, a minimum of three records are required to form a cluster, and the ε distance is set to one. Spatially, if the frequent boarding (or alighting) stops along the recurring routes are adjacent to each other, these stops may be considered as an identical origin (or destination). Therefore, an additional algorithm was used to detect the spatial relationship between multiple routes and applied in the process of DBSCAN clustering, as follows:

Step 1: Randomly retrieve one record that is flagged as unvisited from the sorted trip chain database for an individual smart card. Flag this record as visited and form a cluster for this record.

Step 2: Check the boarding time difference between unvisited records and the last visited record. If the difference is greater than one hour, repeat Step 1.

Step 3: Check the spatial relationship between unvisited records and the last visited record. If a spatial relationship exists (within 200 meters), then this record is included into the cluster formed in Step 1 and flagged as visited.

Step 4: For each cluster, if the number of total records is less than 3, then these records in the cluster are flagged as noise; otherwise, the new cluster is confirmed.

Step 5: Continue to process those unvisited records from Step1 through Step 4 until all the records are flagged as visited.

Step 6: The number of total clusters is the number of typical trip chains per day. The recurring route, boarding/alighting stops and timings can be acquired by counting the most frequent pattern within each cluster.

Take the trip chain data from Table 4-1 as an example. Based on the DBSCAN clustering algorithm, several patterns can be inferred:

This transit rider regularly starts his/her first trip around 7:00 AM, and ends his/her last trip around 6:00PM.

Recurring routes occur in most days of the week. Although the unusual travel pattern is detected on July 8th, it is flagged as a noise by the DBSCAN algorithm.

As previously mentioned, transit riders may take different routes to the same location. The rider took another route, route 350, on July 6th; however, route 350 shares the same stops with route 635. Therefore, two routes are considered as a "common" route, and the shared stops are grouped together.

The routes and stops frequently visited by the same transit rider are demonstrated on a Geographic Information Systems (GIS) map as shown in Figure 4-3. The arrows show the weekday pattern the transit rider follows. It is very likely that this rider takes a home-to-work trip in the morning and then returns to home from his/her workplace in the evening.

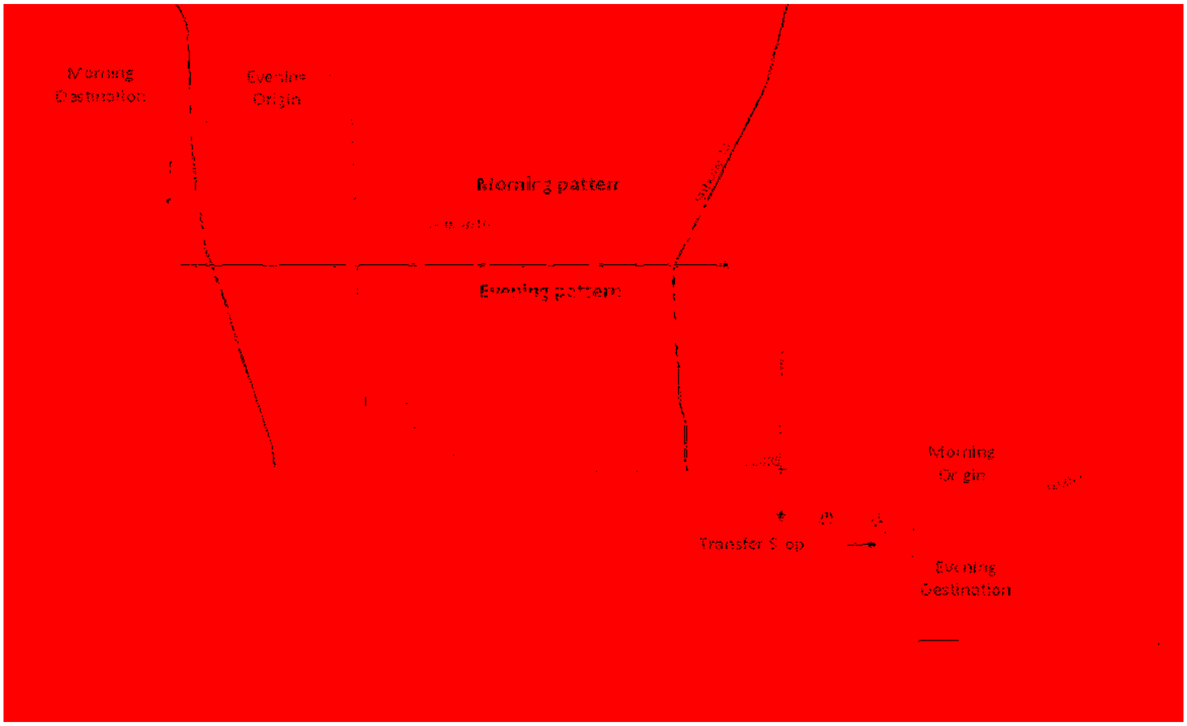


Figure 4-3 Example of a Transit Rider's Travel Pattern

4.3 Regularity Clustering

The historical travel pattern for a particular transit rider can be successfully extracted using the above procedure, but their individual travel pattern regularity is still unknown. As explained earlier, in this context regularity means “how regularly the transit rider travels following the same pattern.” Identifying travel pattern regularity would help transit agencies evaluate the impacts of transit service provision and potential network changes, enabling them to conduct more effective marketing campaigns and measure transit performance (Foote et al., 2001).

Clustering algorithms have been widely used to investigate customer loyalty in the retail and on-line shopping industries (Mauri, 2003; Cheng and Chen, 2009). The same principle can be applied to cluster transit riders with similar travel patterns and place them into different regularity levels based on their temporal and spatial characteristics. Several attributes in the trip chain data were therefore selected as features for clustering as follows:

- Number of travel days

The more days a transit rider travels, the more likely it is that he or she is a frequent transit rider.

- Number of similar First Boarding Time

Boarding time represents a rider’s temporal characteristics. If a rider begins his or her trip at a similar time of day every weekday, then this rider is more likely to be a regular transit rider.

- Number of similar Route Sequence

Route sequence represents a general spatial pattern for a rider. The number of similar route sequences followed during the week may indicate a repetitive travel pattern.

- Number of similar Stop ID Sequence

The Stop ID sequence may contain detailed spatial similarity information. In many cases, two different Stop IDs might be spatially adjacent, which can be identified by GIS buffer processing.

There may be a certain level of correlations between selected features, such as the numbers of similar Route Sequences and Stop ID sequences. However, these correlated features should not be eliminated since there are missing values within the Beijing transit smart card data and introducing a certain level of redundancy into the travel regularity clustering can help improve the algorithm accuracy. Redundant features (e.g. the number of similar stop IDs and the number of similar route sequences) can thus lead to more accurate clustering results.

In order to efficiently and effectively cluster regularity, a suitable clustering algorithm needs to be chosen. The K-Means algorithm is one of the well-known clustering algorithms. This algorithm tries to partition n records into k clusters by minimizing the within-cluster sum of squares. By continuously updating the mean of the record values, each observation is assigned into the cluster with the nearest center until no more observations can be assigned (Forgy, 1965). Although the K-Means algorithm can demonstrate a very high performance and has been applied in many fields, the algorithm suffers from two major intrinsic disadvantages. First, the K-Means algorithm relies on the random initialization of the cluster center and the solution may fall into a local optimum instead of the global optimum as a result of the selection of starting points. If the starting points are far from the true centers of the clusters, the clustering result tends to be locally optimized. Second, the algorithm could require a super-polynomial run time in the worst scenario.

K-Means++, which was proposed by Arthur and Vassilvitskii (2007), addresses the first of these issues by enhancing the initialization process of the traditional K-Means algorithm using a randomized seeding technique to guarantees the optimal solution is obtained. An additional benefit is that the computational complexity of the K-Means++ algorithm is only $O(\log k)$, where k is the number of clusters. More details of the K-Means++ algorithm can be found in Arthur and Vassilvitskii (2007).

To equalize the magnitude and variability of the four input features, variable standardization is conducted before clustering. The range of each variable serves as the divisor to ensure each standardized variable falls between 0 and 1:

$$z = (v - \min(v)) / (\max(v) - \min(v)) \tag{4-1}$$

K-Means++ was therefore chosen to cluster transit riders with similar travel patterns, and each standardized variable can then be incorporated during the travel pattern clustering process.

Five clusters of regularity are used here: Very High (VH), High (H), Medium (M), Low (L), Very Low (VL). The cluster centers can be expressed as:

$$\begin{aligned} c_1 &= (v_{11}, v_{12}, v_{13}, v_{14}) \\ c_2 &= (v_{21}, v_{22}, v_{23}, v_{24}) \\ &\vdots \\ c_5 &= (v_{51}, v_{52}, v_{53}, v_{54}) \end{aligned} \tag{4-2}$$

where v_{ij} represents the j th feature of the generated attributes from trip chain data, and i refers to the i th cluster.

Then, the Euclidean distance between c_i and the zero point is calculated. This distance is defined as the cluster center distance:

$$\begin{aligned} D_1 &= \sqrt{(v_{11}-0)^2+(v_{12}-0)^2+(v_{13}-0)^2+(v_{14}-0)^2} \\ D_2 &= \sqrt{(v_{21}-0)^2+(v_{22}-0)^2+(v_{23}-0)^2+(v_{24}-0)^2} \\ &\vdots \\ D_5 &= \sqrt{(v_{51}-0)^2+(v_{52}-0)^2+(v_{53}-0)^2+(v_{54}-0)^2} \end{aligned} \tag{4-3}$$

Next, D_i is sorted in a descending order. Based on the order, each regularity level is assigned to a cluster. Finally, the corresponding regularity level for each transit rider can be determined by computing and comparing the minimum distance to the center of each cluster.

A preprocessing data cleansing procedure was adopted to eliminate those smart card records with wrong transaction times; for example, a few smart card transactions were recorded as “1900/01/01”. Applying the above data quality control procedure, 37,001 smart cards were randomly selected to test the proposed algorithm.

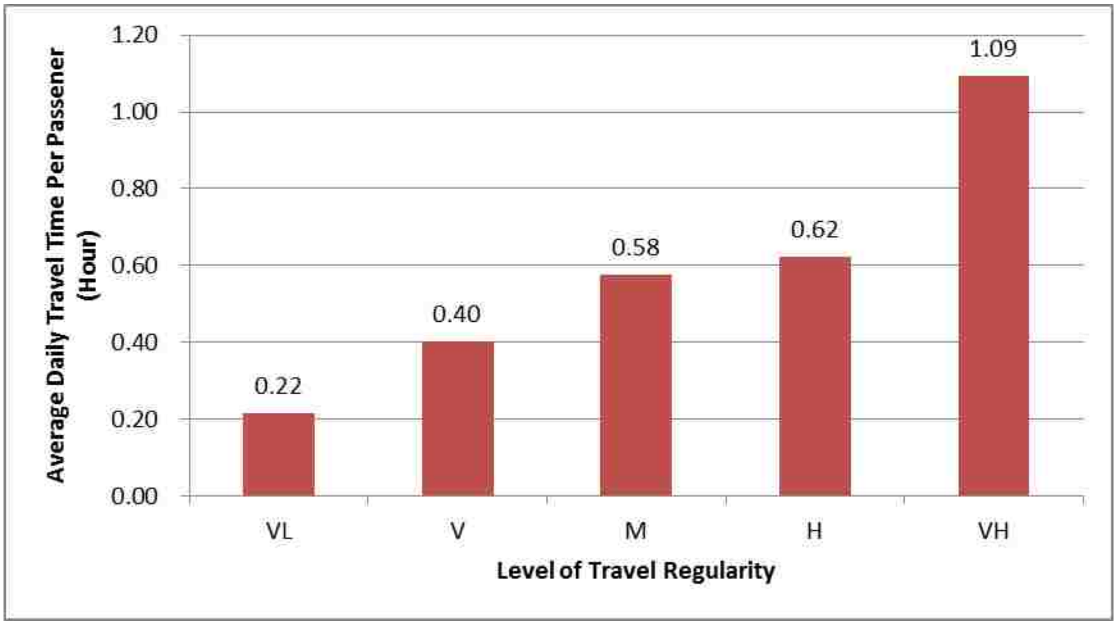
The clustered results are summarized in Table 4-2. If regularity levels of Very High (VH) and High (H) are considered to represent regular transit riders, approximately 41% fall into this category. The clustered results can be used to categorize different transit rider groups for various transit fare options, and provide data support for transit market analyses.

Table 4-2 Summary of Five Clusters

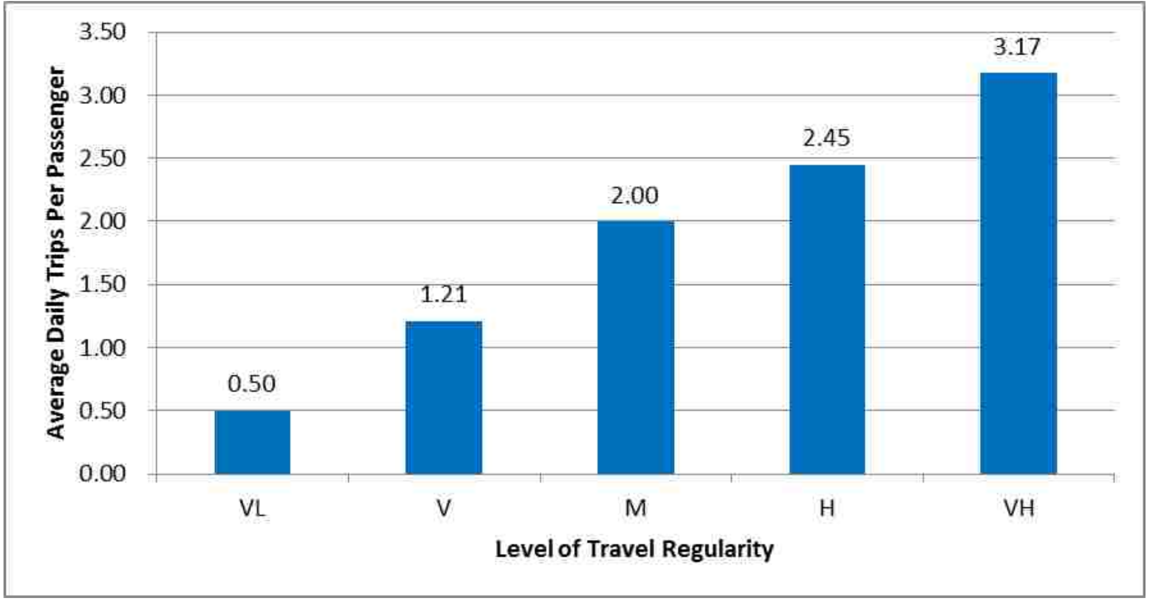
Cluster Center	C_1	C_2	C_3	C_4	C_5
Regularity	VL	L	M	H	VH
Cluster Center Distance	1.28	5.17	10.42	13.52	19.99
Number of Smart Cards	4809	10330	6483	9502	5877
Percentage of total	13.0%	27.9%	17.5%	25.7%	15.9%

Note: VL=Very Low; L=Low; M=Medium; H=High; VH=Very High

Additional individual-level daily trip and travel time information is provided in Figure 4-4. Both average daily trips and average daily travel time for each passenger increased as the corresponding travel regularity became higher. On average, regular transit riders (high regularity and very high regularity) traveled more than twice per day. This is reasonable, because most regular riders take buses at some point during their daily commute.



(a)



(b)

Figure 4-4 (a) Individual-level Average Daily Travel Time for Each Cluster and (b) Individual-level Average Daily Trips for Each Cluster

4.4 Performance Enhancement using Rough Set Theory

In Beijing, more than 16 million smart card transaction data points are generated every day. Processing and clustering such a huge amount of data is not an easy task due to the physical memory constraints of the currently available computer technology. Therefore, the K-Means++ algorithm may not be feasible for this situation without utilizing distributed computing (Cordeiro et al., 2011). To implement and execute the proposed approach in a regular personal computer, an algorithm based on the rough set theory was therefore applied to improve clustering performance. The rough set theory initially proposed by Pawlak (1982) is primarily used to classify vague and uncertain data to help expert systems learn from training datasets and generate meaningful rules for classification. Unlike other commonly used data mining algorithms, rough set-based algorithms do not need any prior information about the data, such as the membership function used in the Fuzzy theory, and the Bayesian prior probability in the Naïve Bayes classifier. Rough set-based algorithms can deal with both continuous and discrete input data, and perform well under circumstances where there is missing or incomplete information. This is because rough set theories depict missing attributes using lower and upper approximations for the incomplete data, defined by probabilities (Grzymala-Busse and Grzymala-Busse, 2007). Consequently, the rough set-based algorithm was deemed appropriate for dealing with the lack of boarding and alighting stop data for the flat-fare buses.

The essence of the rough set-based algorithm is set approximation. Let us define any information system as the form: $\mathrm{A} = (U, A \cup \{d\})$, where U means the non-empty set of objects, also known as universe, A denotes the condition attributes, and d denotes the decision attributes. In our cases, the number of travel days, the number of similar first boarding times, the number of similar route sequences and number of similar stop ID sequences are all condition attributes, and the rider's regularity level is expressed as the decision attribute. The names of the condition attributes and the decision attributes are considered as the universe. The condition attributes and the decision attribute follow a many-to-one relationship. That is, different decision attributes could be sufficiently discerned using only a subset of condition attributes. Therefore, the goal of a rough set-based algorithm is to determine the smallest number of condition attributes to represent the decision attribute. To depict the information uncertainty and vagueness, two important concepts are described as follows. Let $B \subseteq A$ and $X \subseteq U$

$\underline{B}X = \{x \mid [x]_B \subseteq X\}$ is defined as the B-lower approximation of X.

$\overline{B}X = \{x \mid [x]_B \cap X \neq \varnothing\}$ is defined as the B-upper approximation of X.

$BN_B(X) = \overline{B}X - \underline{B}X$ is defined as the B-boundary region of X. If the B-boundary region is not empty, then the set X is considered "rough". (Komorowski, et al., 1999)

Using the above three definitions, we can remove the superfluous attributes and achieve the equivalence classes with the minimum attributes (rules); however, finding the minimum rules is a NP-hard problem and cannot be solved in a polynomial time (Skowron and Rauszer, 1992). Fortunately, many algorithms have been proposed for an optimal solution in an efficient fashion. Wróblewski (1998) developed a fast-rule induction algorithm based on a covering approach. His algorithm has demonstrated its capability in both efficiency and accuracy. Its computational complexity is only $mn\log(n)$, where m is the number of universes and n is the number of attributes.

This rule induction algorithm is used to generate minimum decision rules in our application. Decision results classified by the K-Means++ algorithm are served as training data (as shown in Table 4-3). Then, the rough set theory is applied to extract the hidden classification rules. Example rules are presented as follows:

(1) (Number of the traveling days in (5.75;7.75)) & (Number of the similar boarding time in (7.25;13.25)) => (Regularity Level= High)

(2) (Number of the traveling days in (17.0; Infinity))=> (Regularity Level= Very High)

(3) (Number of the traveling days in (-Infinity;2.0))&(Number of the similar route sequence in (-Infinity;2.5))&(Number of the similar boarding time in (-Infinity;0.5))=> (Regularity Level= Very Low)

The rules determined by the rough set theory can be used to classify each transit rider into different levels of travel pattern regularity. In addition, these rules can be easily implemented and executed in the relational database, e.g. Structured Query Language (SQL) database.

4.5 Comparison of Data Mining Algorithms

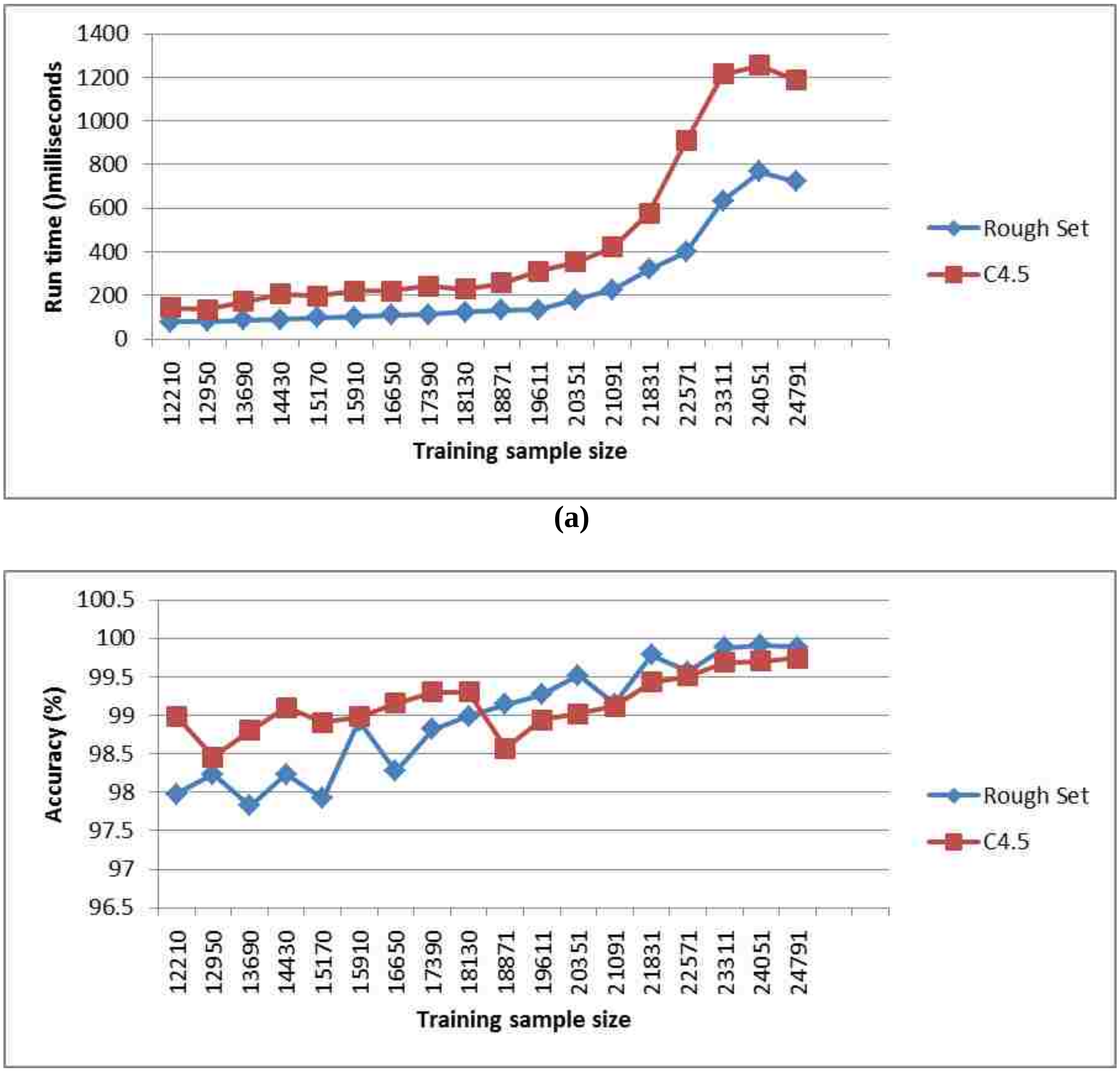
The accuracy and efficiency of proposed rough set-based algorithm were compared with those of several prevailing classification algorithms commonly used in transportation engineering research. namely Naïve Bayes Classifier (Cestnik, 1990), C4.5 Decision Tree (Quinlan, 1993), K-Nearest Neighbor (KNN) (Cover and Hart, 1967) and Three-hidden-layers Neural Network (Rumelhart and McClelland, 1986). The K-Means++ algorithm was adopted as the index algorithm for comparison, with 33% of the clustered transit riders serving as its training dataset. The rough set-based algorithm and the other four classification algorithms were applied in the training dataset to produce the corresponding classifiers. The total sample size was 37001. These classifiers were then used to process the remaining data, and the generated outputs compared to the clustered transit riders obtained using the K-Means++ algorithm to validate the accuracy of each algorithm. The entire dataset was randomly split into 33% training data and 67% test data and each algorithm executed for 10 iterations. All the algorithms were implemented in Java under an environment of a 6-core CPU and an 8 GB RAM desktop computer using the smart card data stored in Microsoft SQL server 2008. Table 4-5 summarizes both the accuracy and run time (the duration taken for an algorithm to execute) statistics of all algorithms.

Table 4-3 Accuracy and Run Time Comparisons among Different Algorithms

Iterations	Rough Set		C4.5		Naïve Bayes		K-NN		Neural Network	
	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)	Accuracy (%)	Time (ms)
1	98.86	59	99.31	116	87.98	64	98.57	802	97.54	120153
2	99.01	54	99.77	119	86.34	66	99.13	798	98.12	119868
3	99.23	69	99.60	123	87.03	65	98.96	793	96.88	123658
4	98.92	64	99.13	118	86.99	70	99.08	826	97.65	130147
5	98.65	59	99.00	127	85.55	68	98.97	757	98.11	121795
6	99.19	53	98.98	133	86.78	71	99.39	788	97.96	116583
7	99.25	60	99.13	130	87.22	69	99.12	809	97.93	130414
8	98.97	51	99.86	121	89.53	69	98.30	825	98.21	125478
9	99.42	56	99.75	109	88.11	65	99.35	786	97.98	123697
10	99.26	63	99.33	111	88.02	72	99.44	779	98.14	130186
Average	99.298	59.8	99.53	113.7	87.649	67.7	99.432	778.4	98.266	123640.5

ms= millisecond

The results show that the proposed rough set-based algorithm clearly outperforms other algorithms in terms of efficiency. A t-test was conducted to evaluate the significance of accuracy the difference in accuracy between the proposed rough set-based algorithm and the other four algorithms. At a 95% confidence level, the proposed algorithm did not significantly differ from the K-NN algorithm but was 10 times faster. In addition, the proposed algorithm outperformed the Naïve Bayes and Neural Network in both algorithm accuracy and efficiency. Although the proposed algorithm slightly underperformed the C4.5 decision tree algorithm in terms of accuracy, it was still twice as fast. As shown in Figure 4-5(a), the rough set-based algorithm demonstrated its strength in efficiency as the size of the training dataset increased. Moreover, Figure 4-5(b) shows that the rough set-based algorithm would outperform the C4.5 decision tree algorithm in terms of accuracy once the size of the training dataset exceeded a certain threshold. This strongly suggests that the proposed rough-set-based algorithm is indeed suitable for handling large datasets of this type.



(a)

(b)

Figure 4-5 Performance Comparisons between the C4.5 and Rough Set-based Algorithms: *(a)* Efficiency comparison, and *(b)* Accuracy comparison

4.6 Discussion

This study opens up interesting new opportunities for leveraging smart card data to create a better understanding of transit riders' behavior and thus potentially improve public transit systems. Specifically, three major potential applications that could benefit from this study can be envisioned, as follows:

- Travel behavior research

In the past few decades, travel demand research has shifted from a trip-based travel approach to an activity-based travel paradigm. Activity-based travel models require a substantial amount of detailed behavioral information for each traveler extending over a relatively long period. Traditionally, this type of behavioral data has been collected using travel diaries and travel surveys, requiring an immense amount of resources to process and construct sequences of spatiotemporal activities for each traveler (Schlich and Axhausen, 2003).

The individual-level travel pattern mining algorithms developed for this study offer an alternative and novel approach for measuring the similarity and variability of transit riders through an examination of their multi-day smart card transactions. This will greatly facilitate travel behavior modeling development.

- Transit market analysis

As with other domains such as eGrocery shopping, transit agencies generally aim to develop a range of different market strategies to satisfy their passengers (Strathman et al., 2008). One typical application of transit market analysis is market segmentation (Zhou et al., 2004). Market segmentation techniques divide the entire market into several distinct segments consisting of groups of transit riders who share similar preferences and attitudes. Based on the transit rider groups identified here using the proposed travel regularity clustering algorithm, transit agencies can better allocate their limited resources to each segment to maintain and attract ridership. For example, various transit fare option can be provided that are specifically tailored for each group of transit riders. Key factors that influence transit ridership can also be identified by integrating each market segment with transit riders' socio-demographic attributes (Krizek and El-Geneidy, 2007). For instance, most regular transit riders are commuters who do not own private cars and thus tend to be very sensitive to service reliability. In this case, improving transit service reliability (by, for example, shortening headway and providing real-time information) could be an effective measure to retain this group of transit riders.

- Transit OD estimation

Another potential use of the proposed travel pattern and travel regularity mining algorithms is to improve the accuracy of the transit OD estimation method. Each transit rider's repetitive historical routes and stops can be used as prior information for passenger alighting stop inference.

4.7 Conclusion

This chapter proposes a series of efficient and effective data-mining approaches with which to model transit riders' travel patterns using the smart card data of the type collected in Beijing, China. The DBSCAN algorithm was utilized to successfully detect each transit rider’s historical travel pattern using the identified trip chains. The K-Means++ clustering algorithm and the rough set were then jointly applied to classify the travel pattern regularities. The performance of the resulting rough-set-based algorithm was compared with four other classification algorithms: the Naïve Bayes Classifier, C4.5 Decision Tree, K-Nearest Neighbor (KNN) and three-hidden-layers Neural Network. The results indicated that the proposed rough-set-based algorithm outperformed all the other data mining algorithms in terms of accuracy and efficiency.

The contribution of this study is two-fold: First, a data mining approach has been proposed that is capable of identifying travel patterns for individual transit riders using a large smart card dataset. The second contribution is that the regularity levels for the data can also be successfully classified by the approach proposed here. The travel patterns and regularity levels of their customers are important information for transportation researchers seeking to understand day-to-day urban travel behavior variability and facilitate activity-based travel demand model development.

Individual travel patterns and pattern regularity also offer substantial benefits for transit agencies working to improve their transit service with the assistance of transit market analysis. Another potential application of this research is to estimate an individual transit rider’s origin and destination using that rider’s historical travel pattern. In terms of future work, the proposed method must now be compared with other traditional travel behavior data collection methods, such as survey studies, focus group discussions and travel diaries, in order to improve the algorithm accuracy. It would also be interesting to integrate the passenger travel pattern information obtained through this study with map-based transportation systems (Ma et al., 2011) to monitor and visualize transit performance.

Chapter 5 Transit Passenger Destination Estimation

For Beijing’s flat-fare based AFC system, transit riders are not required to tap their smart card during checking out, and consequently, the alighting stop for each transit rider is missing. Therefore, it is of critical need to estimate the alighting stop for each SC transaction. In this chapter, the alighting stop for each transit smart card holder will be inferred. Three major analytical methods are adopted: (1) spatiotemporal transfer activity identification (2) daily trip chain analysis (3) historical travel pattern integration. Smart card transactions from distance-based fare buses with both boarding stop and alighting stop information will be used to validate the proposed alighting stop inference algorithm.

5.1 Methodology

5.1.1 Spatiotemporal Transfer Activity Identification

Based on the previous chapter, the passenger boarding stop for those flat fare buses can be either estimated by integrating with GPS data or by applying the Bayesian decision tree algorithm. Transfer activities between various bus lines and subway lines are quite common in Beijing transit networks. This passenger transfer information can be utilized for passenger alighting stop estimation with the following assumption: the passenger alighting stop of the current trip is spatially and temporally adjacent to the boarding stop of the next trip. To be specific, there are three types of transfer activities for the flat-fare based route:

(1) A transit passenger can transfer from a flat-fare route to a subway line.

(2) A transit passenger can transfer from a flat-fare route to a distance-based fare route.

(3) A transit passenger can transfer from a flat-fare route to another flat-fare fare route.

For scenario (1), passengers are required to tap their smart cards on the card reader when they enter and exit through the gateway. Both boarding and alighting information (location and time) are stored by the card reader. For scenario (2), two smart card readers are installed besides the front and rear doors of distance-based fare buses, and they can record each passenger’s boarding stop and alighting stop respectively, however, the boarding time information is not able to be stored when each passenger gets on the bus. Fortunately, other passengers’ alighting time can be a good substitute of the missing boarding time since the time difference between alighting activity and boarding activity is marginal. For scenarios (3), the boarding time for each flat-based route is accurately captured by the smart card reader, and most of passengers’ boarding stops can be inferred within an acceptable range of error using data mining and data fusion approaches.

To connect two distinct transit routes as a linked trip for a passenger, one assumption is that the passenger should board the next route within 60 minutes after boarding the next route for transferring purpose. The 60-minute threshold is based on the result from 2010 Beijing transport survey (Beijing Transportation Research Center, 2012), where the average transit trip time (in-vehicle time and waiting time) per passenger is 65.4 minutes. In addition, the boarding stop from the current route should be within a 200-meter vicinity of the alighting stop from the previous route when a passenger transfers between these two routes. A passenger tends to choose the stop within minimum walking distance, and therefore this stop will be the most probable alighting stop for this passenger.

5.1.2 Daily Trip Chain Analysis

The above transfer activity analysis can estimate most passengers' alighting stops for flat-fare routes. There are a certain amount of smart card transactions without any associated transfer activities such as the last transaction of a particular day. One assumption can be made for alighting stop inference: passengers will return to his/her first boarding stop at the end of a day, and this first boarding stop should be the origin of his/her first trip of the same day. This assumption can be better illustrated by figure 5-1. The passenger started his/her first trip from the origin (blue stop), and transferred from the green stop to the destination i (squared stop) along the second route. At the end of the day, this passenger took the returning route and alighted at stop k (squared stop) in the vicinity of his/her origin (blue stop). As shown in figure 5-1, the last trip of the passenger is likely a trip from work to home. In this case, alighting stop k in the second route can be estimated using the origin (blue stop) of the returning route. Similarly, the first trip is a home to work trip, and thus, the alighting stop i can be substituted by the origin (blue stop) of the first route. It is worth noting that it is not necessary for a passenger taking the identical route for commuting, and this is to say, each passenger could have several different transit routes from his/her home to his/her workplace. Therefore, further spatial analysis should be conducted to justify whether two different routes share several geospatially similar stops.

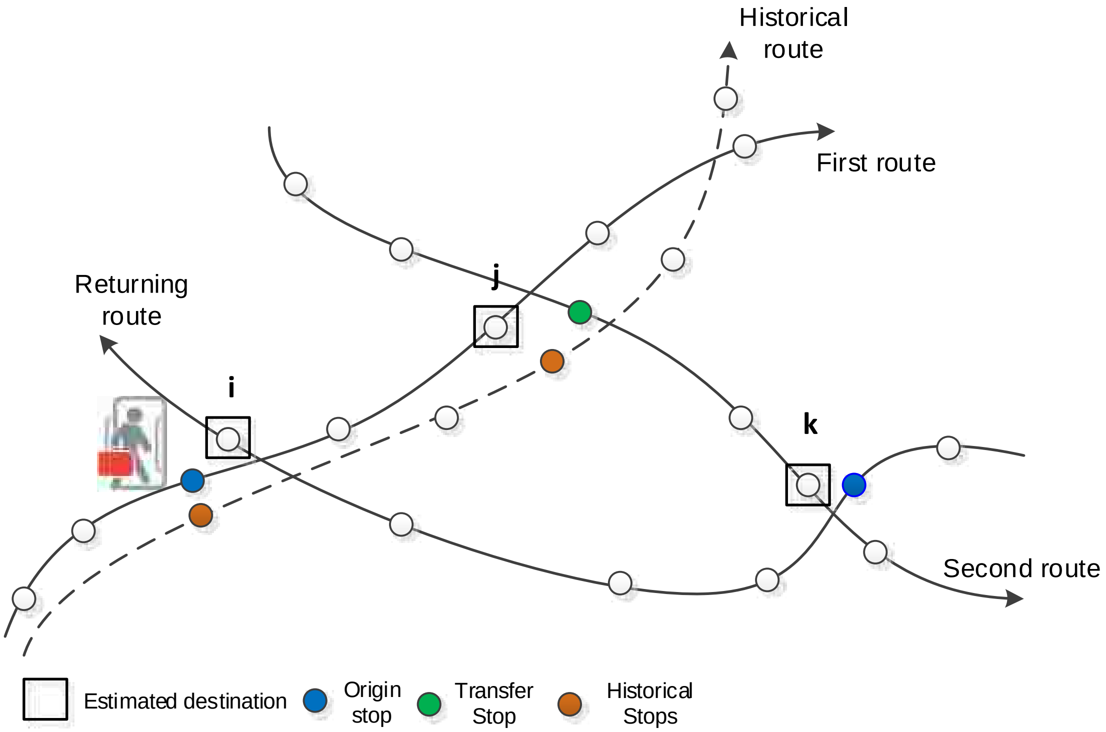


Figure 5-1 Alighting Stop Inference Algorithm Example

5.1.3 Historical Travel Pattern Integration

In our study, multi-day smart card transactions for each particular transit rider are also taken into account for passenger destination estimation. Using the historical individual level travel pattern, we are able to infer the most probable destination for the same transit rider on a different day. The underlying assumption is that the transit rider with high travel regularity is likely to follow his/her historical travel pattern to repeat his/her future trips. If a transit rider's historical travel pattern has been successfully detected using aforementioned data mining approach, and this transit rider's travel regularity is simultaneously classified as "high", then the missing alighting stop of the current trip for this transit rider can be substituted by the known historical alighting stop. The dashed line in Figure 5-1 represents the historical route for the transit rider identified from multi-day smart card transactions data, and this route is a distance-based fare route with both known boarding and alighting stops. In addition, the most frequent stop ID sequence can be also identified using the DBSCAN clustering algorithm in Chapter 4. Assuming this regular passenger continued to follow his/her previous routes as the current travel pattern, the most spatially nearest stop to the historical alighting stop (orange stop) can be identified as the alighting stop along the first route.

Based on the above three primary approaches, the fundamental flow chart to infer individual passenger's alighting stops can be drawn as follows:

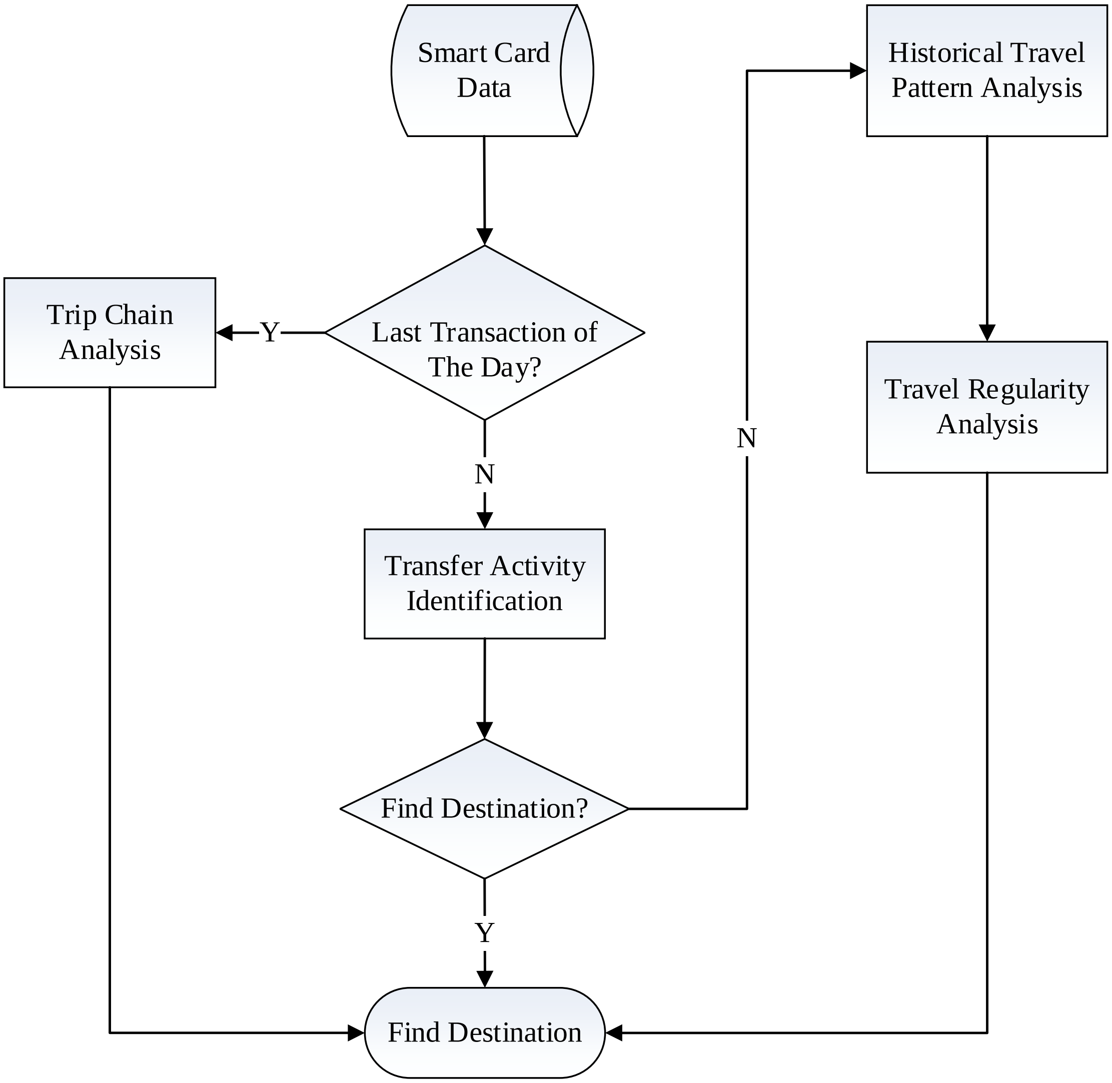

Smart Card Data
Historical Travel Pattern Analysis
Trip Chain Analysis
Y
Last Transaction of The Day?
N
Travel Regularity Analysis
N
Transfer Activity Identification
Find Destination?
Y
Find Destination

Figure 5-2 Transit Passenger Destination Estimation Flow Chart

5.2 Validation

Alighting stop information is absent for flat-fare routes, however distance-based fare routes contain both alighting and boarding stop locations. These known alighting stop are considered as “ground-truth” data, and will be compared with the inferred alighting stops using the proposed destination estimation algorithm for validation purposes. Two distance-based fare routes (route 753 and route 967) were selected. A total of 27343 distinct smart cards were extracted from route 753 on July, 5th, 2011. In addition, one-weekday (June, 27th, 2011 to July, 1st, 2011) smart card transactions from the 27343 smart card holders were collected to generate each passenger's typical travel pattern and travel regularity. Based on the travel pattern and travel regularity data mining approach discussed in Chapter 4, the statistics for travel regularity were generated and the results are summarized in Table 5-1:

Table 5-1 Statistics of Travel Regularity for Route 753 on July, 5th, 2011

Number of Smart Cards	Percentage of Total	Regularity Level
5571	20.4%	VL
7848	28.7%	L
6750	24.7%	M
3202	11.7%	H
3972	14.5%	VH

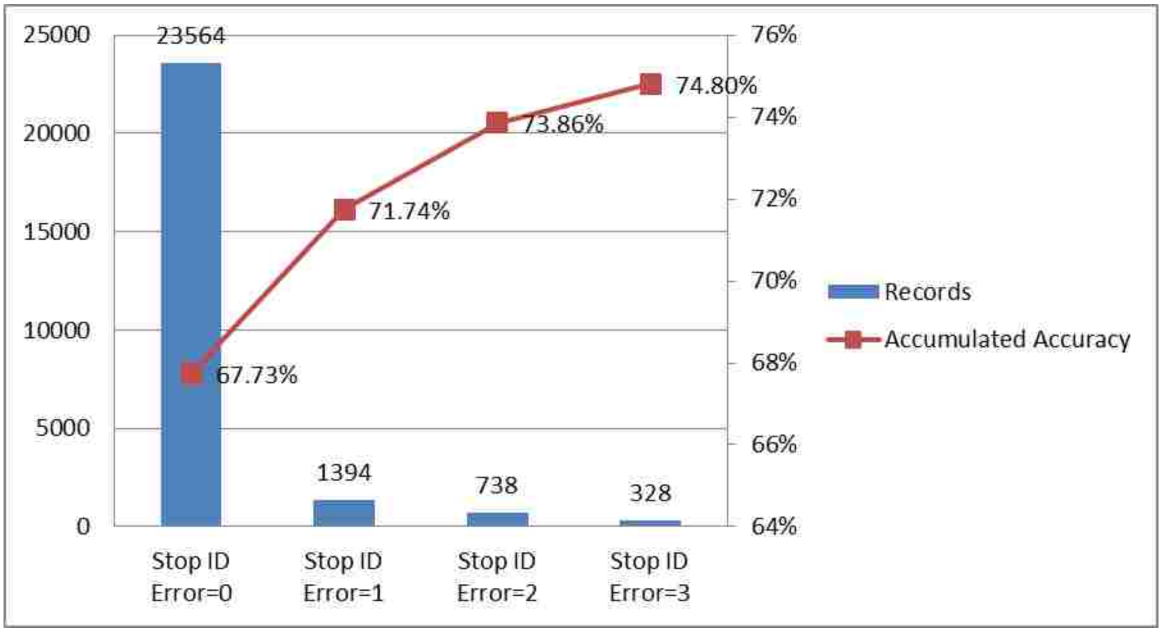
The inferred individual travel pattern information is then incorporated into passenger destination estimation algorithm. The total sample size of smart card transactions is 34790. The results are shown in Table 5-2 and Figure 5-3:

Table 5-2 Results of Alighting Stop Estimation Algorithm for Route 753

Route 753	Number of records	Accumulated percentage in inferred records	Accumulated percentage in total records
Stop ID error<1	23564	86.2%	67.7%
Stop ID error<2	24958	91.3%	72.7%

Stop ID error<3	25696	94.0%	73.9%
Stop ID error<4	26024	95.2%	74.8%
Total	27337	N/A	78.6%

Figure 5-3 Alighting Stop Estimation Algorithm Accuracy for Route 753



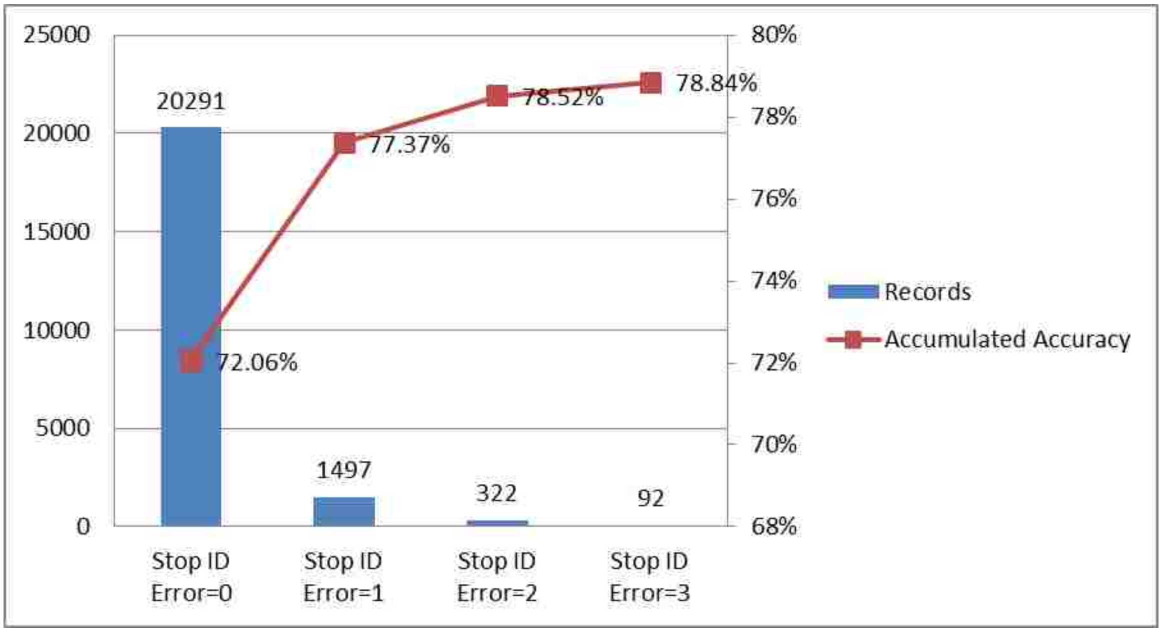
Alighting stops for 78.6% of the smart card transactions can be estimated by the proposed algorithm. Among these inferred stops, 86.2% of them match with the actual alighting stops, and more than 95% stops fall within three-stop distance away from the actual alighting stops.

Similarly, a total of 28160 smart card transactions from route 967 on July, 5th, 2011 were also tested. 81.8% of the total alighting stops can be successfully estimated, which contain 88.1% perfectly matched records, and more than 96% records whose error ranges are within three stops. The tested result is displayed in Table 5-3 and Figure 5-4.

Table 5-3 Results of Alighting Stop Estimation Algorithm for Route 967

Route 967	Number of records	Accumulated percentage in inferred records	Accumulated percentage in total records
Stop ID error<1	20291	88.1%	72.1%
Stop ID error<2	21788	94.6%	77.4%
Stop ID error<3	22110	96.0%	78.5%
Stop ID error<4	22202	96.4%	78.8%
Total	23032	N/A	81.8%

Figure 5-4 Alighting Stop Estimation Algorithm Accuracy for Route 967



The alighting stop accuracy of route 967 is slightly higher than the one of route 753. This is probably due to the characteristics of transit riders. Table 5-4 summarizes the travel regularity in route 967. If a passenger is classified as high or very high travel regularity, then this passenger is considered as a regular transit rider. More than 43% passengers from route 967 are categorized as regular riders, while there are 26% passengers from route 753 as regular transit riders.

Table 5-4 Statistics of Travel Regularity for Route 967 on July, 5th, 2011

Number of Smart Cards	Percentage of Total	Regularity Level
4101	17.8%	VL
4903	21.3%	L
4010	17.4%	M
6458	28.0%	H
3569	15.4%	VH

This is not surprising. As shown in Figure 5-5, red line represents route 967, and blue line represents route 753. Route 967 goes through the university area (black circle area), where a variety of universities are located. The passengers in route 967 are likely composed of college students. They are probably frequent transit riders for commuting to universities, and this leads to a higher percentage of regular transit riders than another route 753. Therefore, higher alighting stop estimation accuracy can be achieved for route 967.

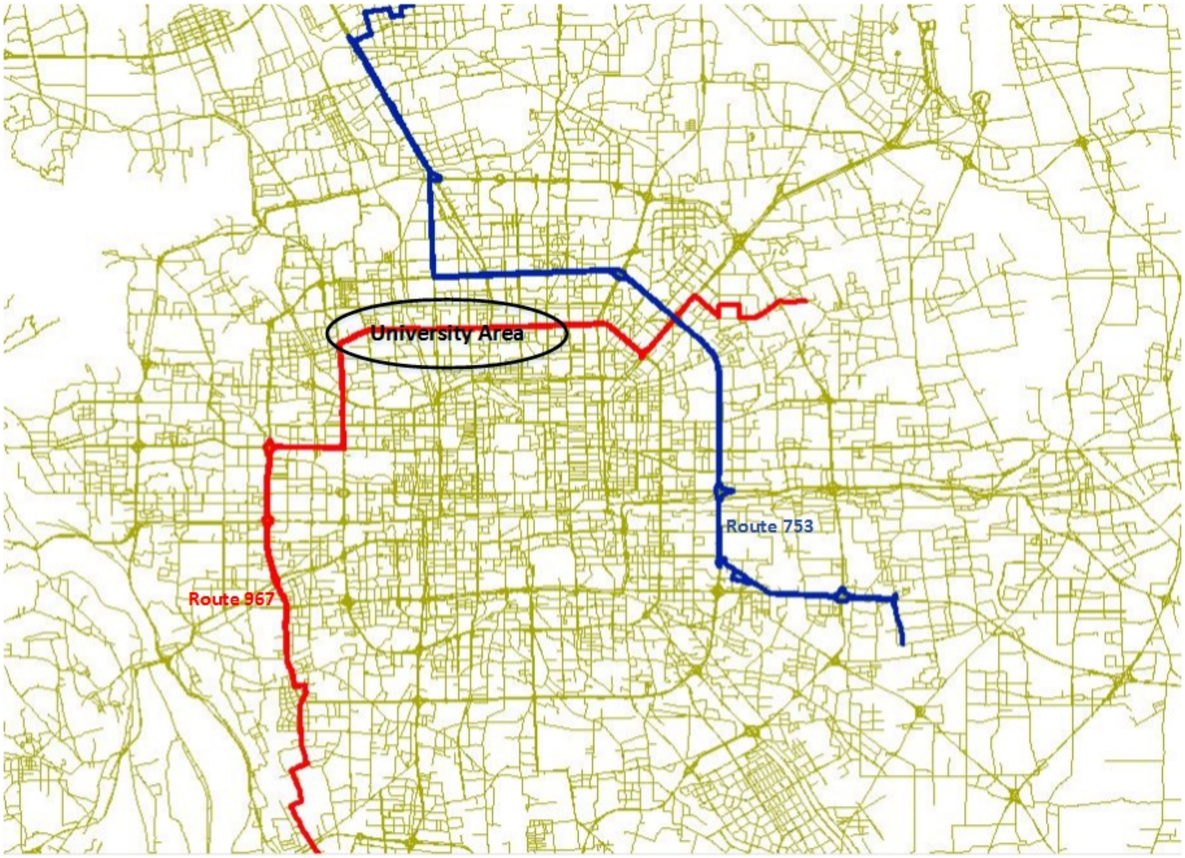


Figure 5-5 Route 967 and Route 753 Spatial Distribution

Both the passenger origin inference algorithm and destination estimation algorithm were implemented in Microsoft Visual C# as displayed in Figure 5-5, and the inferred OD data were stored in Microsoft SQL server 2008.

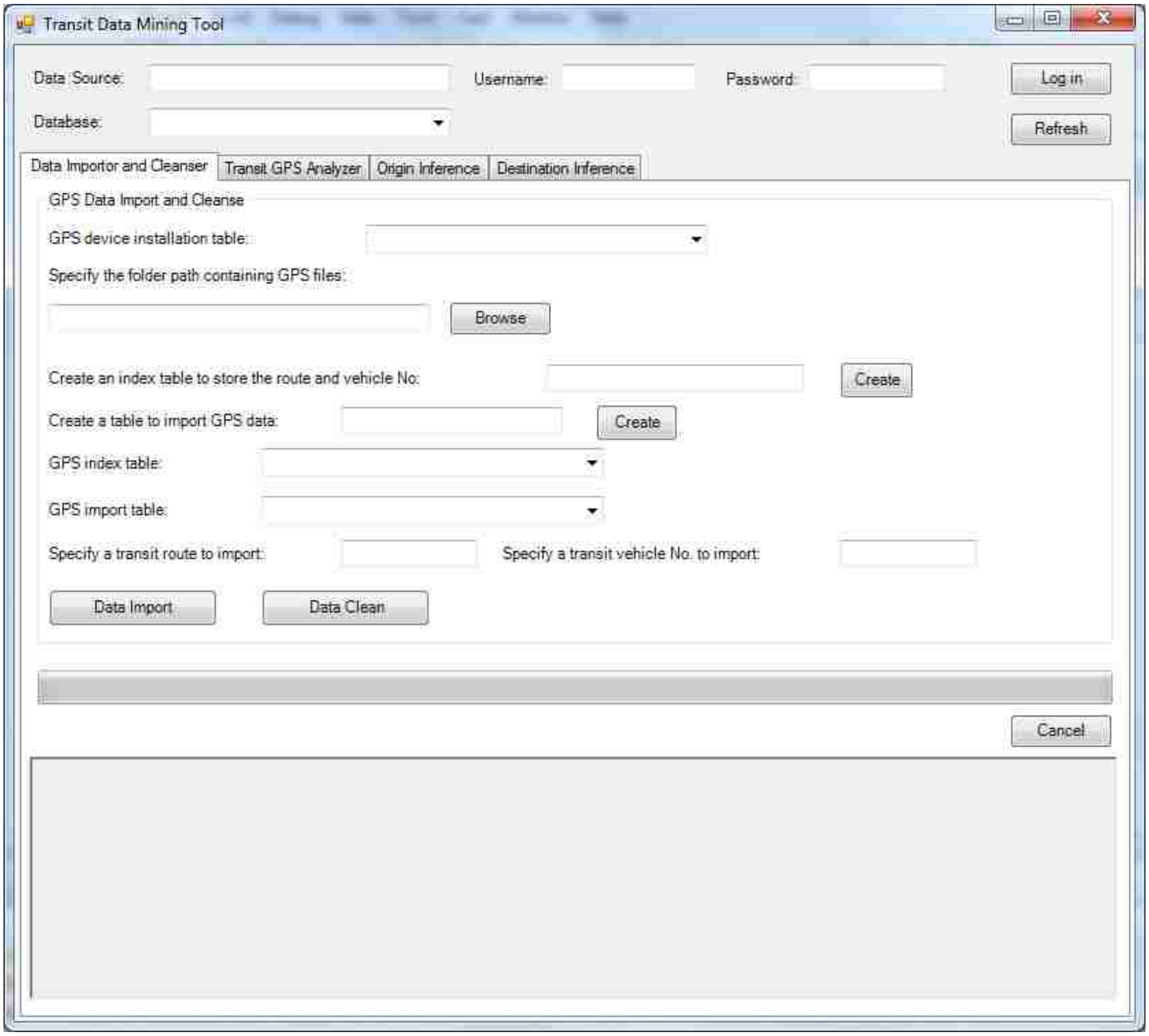

Transit Data Mining Tool
Data Source:
Username:
Password:
Log in
Database:
Refresh
Data Importor and Cleanser
Transit GPS Analyzer
Origin Inference
Destination Inference
GPS Data Import and Cleanse
GPS device installation table:
Specify the folder path containing GPS files:
Browse
Create an index table to store the route and vehicle No:
Create
Create a table to import GPS data:
Create
GPS index table:
GPS import table:
Specify a transit route to import:
Specify a transit vehicle No. to import:
Data Import
Data Clean
Cancel

Figure 5-6 Beijing Transit Origin and Destination Estimation Software

Chapter 6 Transit Performance Monitoring and Visualization

Two research hot spots are identified and should be address in this section: (1) Develop a series of effective performance indicators to better quantify the public transit performance using the individual-level passenger OD and trip information; (2) Develop a robust and efficient visualization platform to convey the transit performance measures to both transit agencies and transit riders. To achieve these goals, several key transit performance indicators are firstly introduced, and the concept of eScience is then further elaborated with a prototype system named as DRIVE Net. DRIVE Net system can be leveraged to calculate and present the performance of transit system based on massive smart card data on different scales. This leads to the development of TransitNet, an E-science based platform to quantify and visualize transit performance measures in a large-scale network. The architecture of TransitNet system will be detailed in the second part of this chapter.

6.1 Transit Performance Measures

In this study, several transit performance indicators are proposed depending on different levels of analytical scopes. The definitions of transit performance measures follow a hierarchical structure from a network-level speed map to a stop-level headway analysis. Each level of transit performance indicators is articulated in the following content.

6.1.1 Network-level Transit Speed

Network-wide transit travel speed is of particular interest for transit agencies and travelers. Using the real-time transit travel speed data, transit path finding algorithm based on the minimum travel time can be readily implemented for transit rider's trip planning. Moreover, transit agencies are also able to identify congestion bottlenecks by observing those low-speed routes, and examining how congestion propagates through the transit network. Consequently, corresponding countermeasures (e.g. transit routing optimization) can be further adopted by transit agencies to alleviate the congestions. However, transit speed cannot be directly measured using the prevailing inductive loop detectors. Especially considering the passenger boarding and alighting activity at a particular bus stop, estimating transit speed may require external data to accomplish. Smart card transaction data contain both temporal information (transaction time) and spatial information (bus route and inferred OD information), and can be used to estimate the travel time between stops. The travel time between two adjacent stops includes both running time and dwell time (Vuchic, 2005). Dwell time is the duration of a bus standing at a stop for the purpose of boarding and alighting passengers, and is highly related to the number of waiting passengers. Actual transit speed calculation should consider both general traffic condition and passenger dwell activity (Vuchic, 2005), and then, the stop-to-stop average speed will be calculated by using the network distance between adjacent stops divided by the stop-to-stop travel time. In the Beijing transit system, passengers are required to simultaneously board a bus through the front door, and get off the bus through the rear door. In this case, the stop-to-stop travel time can be approximately measured as the first passenger's boarding time difference between stop i and stop j.

Transit speed calculations can be undertaken for both flat-fare buses and distance-based fare buses in the entire transit network. Potential applications of the network-wide transit speed include dynamic routing for transit drivers, transit network optimization for transit agencies, and transit congestion diagnosis for transportation researchers.

6.1.2 Route-level Transit Travel Time Reliability

Travel time reliability represents the consistency of travel time of a repeated trip, and is more important than average travel times since travelers tend to remember unexpected delays (Lyman and Bertini, 2008). Providing travel time reliability information can help traveler to better manage their time and reduce traffic congestion. This is particularly true in public transportation. Transit travel time reliability influences transit service attractiveness and efficiency, and relates to on-time performance and headway deviation for customer satisfaction issues (Kittelson&Associates, 2003). Transit agencies are actively seeking solutions to retain their current ridership and attract more transit riders by mitigating the unexpected delay by for example, applying bus-only lanes. From the perspective of a passenger, passenger travel time is composed of both in-vehicle travel time and waiting time (Vuchic, 2005). Route-level transit travel time reliability indicators can be used to measure the variance of in-vehicle travel time. There are a variety of effective methods to quantify travel time reliability recommended by the Federal Highway Administration (FHWA, 2006):

- 90th or 95th percentile travel time

This indicator depicts the travel time during the heaviest traffic conditions.

- Buffer index

Buffer index measures the extra time a traveler needs to spend in addition to his/her average travel time to ensure on-time arrival for 95% percent of the trips. The extra time can be defined as the time difference between 95^{th} percentile travel time and average travel time, and Buffer time is then calculated as the ratio of the extra time and average travel time.

- Planning time index

Planning time index can be calculated as the 95^{th} percentile travel time divided by free-flow travel time. Different from the buffer index, the planning time index represents the necessary total time to ensure on-time arrivals for 95% percent of the trips.

In this study, 95% travel time and buffer index are adopted to measure the transit travel time for a particular route. In addition, the segment-level travel time reliability indicator of each transit route is also calculated for the further evaluation.

6.1.3 Stop-level Ridership and Headway Variance

Ridership

Stop-level ridership refers to the number of boarding or alighting passengers at each stop along a route. Ridership plays a significant role in monitoring transit service and accessing financial gain for transit agencies. Transit operators can use stop-level ridership to identify popular stops with a high number of boarding passengers, and correspondingly adjust transit schedules to achieve better service quality. Decision makers in transit agencies can evaluate the effectiveness of new fare policies and how transit ridership responds to the fare changes.

The total number of alighting and boarding passengers determines the demand for each route and can potentially impact transit agencies' marketing and operational strategies. For example, if a particular route is found to be heavily utilized, then, increasing the number of daily transit vehicles or shortening the headway could better accommodate more transit riders.

Headway Variance

Headway is a key factor to measure transit service reliability, and is defined as the time difference between two consecutive buses at arriving at a particular stop for the same route (Liao and Liu, 2010). Neither too long headway nor short headway is desired by transit agencies and passengers. Too short headway may result in the bus bunching. In this situation, if one bus arrives at a particular stop late due to traffic congestion, the following bus will likely arrives at the same stop after a short time period as well. Consequently, the first bus is full of passengers, while the second bus is almost empty. This leads to inefficient transit usage, and thus should be avoided by transit agencies. Similarly, long headway can incur long waiting time for passengers. As demonstrated by Mohring et al. (1987), passengers value waiting time two to three times more important than in-vehicle travel time. Irregular headway may reduce the attractiveness of public transit. It is worth noting that the headways at different stops may fluctuate due to traffic signals and variable quantities of boarding passengers. Therefore, it is necessary to conduct a stop-level headway analysis for a transit scheduler to adjust the headways in the middle of the route.

Smart card data can assist better assessing the headways. In this study, the stop-level headway can be calculated as the time difference of the first passenger's boarding times for the two consecutive buses, and presented as a histogram for visualization purposes.

6.2 Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net)

Developing the aforementioned transit performance indicators requires interacting with transit network elements (i.e. stops and routes). A transit network can be considered as a dynamic transportation network with spatial and temporal transit features (Huang and Peng, 2008). Because of the complexity in transit network, integrating roadway geometric attributes with transit operational data becomes challenging. According to Transit Cooperative Research Program's report (Furth et al, 2006), transit agencies is of particular interest to conduct route-independent demand and service reliability analyses. However, achieving such a goal is not a straightforward task, which requires a comprehensive GIS model and sufficient computing power to link transit stops and segments, and integrate the transit-related passive data (i.e. smart cart data and GPS data).

The advancements of information technology and mobile application have revolutionized the way of information gathering and dissemination, and a large amount of location-aware data is becoming affordable and ubiquitous. This is especially true in public transportation. There is a strong need to develop a robust web-based platform to increase the exchangeability and usability of intensive transit data. To overcome the shortcomings of traditional static GIS applications for transportation, a web-based E-Science framework has been proposed. This framework is named as DRIVE Net, and the following contents focus on the detailed information of this system.

6.2.1 Motivation

Over the past decades, transportation research has been mathematical-equation driven and relied on scarce data to develop mathematical models and traffic theory (e.g., Chiu and Mirchandani, 2008, Murray-Tuite and Mahmassani, 2005). For example, many well-known theoretical models were developed based on a small portion of data but they have been widely used in practice (May, 1990; Pipes, 1967). When the model development is extended to the network level, data availability may reduce further or simply disappear and these theoretical models can often be verified only by simulation data (e.g., Smith el al., 2008; Haghani et al., 2008). Options did not seem to exist although researchers knew that the simulation and mathematical models only can capture some of the “facts,” many contributors, especially human factors, are not easily reflected in the simulation results.

With the advance of data collection technologies and their deployments in Intelligent Transportation Systems (ITS), transportation data availability has been increasing tremendously over the past years. As a future type of traffic management system, IntelliDrive^{TM} (http://www.intellidriveusa.org/) is also quickly gaining in popularity and increasingly deployed. Since IntelliDrive^{TM} enables vehicle to vehicle and vehicle to infrastructure communications on a frequent basis, traffic data are expected to explode in the years to come. Therefore, data-driven or data-based research shall expand and play an increasingly important role in the near future. The rich data sets will enable validation of previously developed transportation theories and boost scientific discoveries on transportation planning, system operations, and travel behaviors.

Most of the previous web-based Archived Data User Services (ADUS) systems and Advanced Traveler Information Systems (ATIS) are primarily based on a single data source and serves as a traditional online data or online traffic information provider. Despite of the needs from various transportation-related agencies for online systems to share and analyze transportation relevant data, few such systems were developed with the functions of data format standardization, regional map-based data visualization, and interactive online traffic analysis, with consideration of the interactions between heterogeneous data. For example, the impacts of freeway incidents on arterials and freeways were not covered in previous research due to the lack of an explicit architecture to bridge the gaps between heterogeneous data from multiple transportation agencies.

The goal of this study is to develop an e-science of transportation platform for data sharing, visualization, modeling, and analysis. The term “e-Science” was created by Dr. John Taylor of the UK Office of Science and Technology in 1999 (Hey and Trefethen, 2002). E-science refers to computationally intensive science that needs to process immense data sets using highly distributed computational resources connected by the Internet. E-science approaches have a great potential to solve some tough transportation issues. However, it has been a slow process for the transportation communities to accept this new concept. The new platform intends to take advantage of e-science developments for data-driven transportation research and applications.

The new platform is named Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net). DRIVE Net is expected to remove data accessibility barriers, allow easy access of real-time regional traffic information, facilitate data sharing and visualization, enable online data analysis for scientific discovery and decision support, and offer opportunities for early stage e-science of transportation investigations. This platform will benefit not only regular road users but also transportation practitioners and researchers. Compared with previous systems, DRIVE Net is not a simple data visualization and archiving system. DRIVE Net not only enables the connections and interoperability among the heterogeneous data sets but also serves as a data-rich visual platform to facilitate scientific discoveries and educational enrichments in the areas of transportation engineering and planning, environmental engineering, and public health science. The design for DRIVE Net clearly considered the need to support future endeavors in e-science of transportation.

6.2.2 System Architecture

The design of DRIVE Net is critical for its future performance and scalability. The current design shown in Figure 6-1 reflects our current understanding of the platform and future expectations to DRIVE Net. We intend to make it open architecture and open source so that system design can be continuously improved with expansion of system capabilities. The current system architecture is primarily composed of three parts: heterogeneous data sources from different agencies, data warehouse in the STAR Lab, and web services running on the DRIVE Net system server.

Data Warehousing and Retrieval

The data warehouse is responsible for data archiving with multiple data retrieval functions supported by the DRIVE Net system. Data retrieval is challenging because every agency has its internal policy and security concerns. Relying on one single uniform data retrieval method for each agency may be infeasible and inapplicable. Moreover, data archiving formats in all agencies vary, even within the same agency; data may still follow different patterns. Standardization of data formats is highly beneficial for transportation agencies and data users. However, few guidelines have been developed for data exchange and standardization. Some standards, e.g. National Transportation Communications for ITS Protocol (NTCIP) (NCTIP, 1998), only focus on data communication standardization, not the data exchange and storage formats. Hence, four data retrieval methods are proposed and currently used for data retrieval in DRIVE Net. These methods and data examples are listed as follows:

1) Traditional flat file exchange: Flat files are the most common data exchange format commonly used among all agencies. In this way, data quantity, property and privacy can be carefully controlled. However, this method is also less efficient and more time-consuming. Once these files are retrieved through the physical media, e.g. CD-ROM, or e-mails, these files can be uploaded into the database through the DRIVE Net website or using the Structured Query Language (SQL) import function. Below are two examples of data obtained through flat file exchange*:*

- *Washington Incident Tracking System* (*WITS):* Most incidents happening in Washington major freeways and state highways are logged in the WITS database in Washington State Department of Transportation (WSDOT). The WITS datasets are disseminated in flat files (EXCEL) and imported into the DRIVE Net incident database. Detailed incident information, such as incident geospatial location, notification time, clearance time, is stored in the incident database.

- *Highway Safety Information System (HSIS)* (Council and Mohamedshah, 2009): Upon users' request, the HSIS provides different types of data for Washington State highways, including the accident, roadway inventory, traffic volume, curve and grade and interchange/ramp data. The data are stored in the DRIVE Net accident database.

2) Passive data retrieval: DRIVE Net is equipped with customized C# or Java computer programs that are scheduled to fetch the remote data in a predefined interval via

File Transfer Protocol (FTP), Hypertext Transfer Protocol (HTTP) or Simple Object Access Protocol (SOAP). This method is considered the most convenient and efficient way to periodically retrieve data from remote servers. The data can be imported into the database following the schema design by the research team at the Smart Transportation Applications and Research Laboratory (STAR Lab) of the University of Washington (UW). Examples:

- *Freeway loop sensor data:* The Washington State Department of Transportation (WSDOT) operates more than 7000 Inductive Loop Detectors (ILD) along freeways in Washington State (Wang et al., 2009). WSDOT shares its 20-second aggregated single loop data via an FTP website. The data are automatically fetched by the data download module every 20 seconds and stored in the DRIVE Net freeway database.

- *Arterial data:* The City of Bellevue, WA has more than 500 advance loop detectors at more than 177 signalized intersections (Wu et al., 2009; Wu et al., 2007). The controllers at these intersections send the cycle-by-cycle real-time traffic data (e.g. volume, occupancy and timing plans) back to a FTP server in City’s Traffic Monitoring Center (TMC) and stored as Comma-Separated-Value (CSV) file every minute. DRIVE Net has been fetching the CSV file and importing the data into the DRIVE Net arterial database since 2007.

- *Trucking Data:* Global Positioning Systems (GPS) used by trucking companies are a source of truck probe data for freight performance measurement. WSDOT, UW and the Washington Trucking Associations (WTA) have partnered to collect and analyze GPS truck data from commercial in-vehicle fleet management systems used in the central Puget Sound region (Ma et al., 2011). Data are being collected from three vendors, with various resolutions, ranging from one to 15 minute frequency. DRIVE Net automatically fetches and imports theses data into DRIVE Net truck database via FTP.

3) Active data retrieval: Some agencies may have internet security concern and limited public access. The STAR Lab provides a satellite server with hardware, software, and data processing tools pre-installed. Using a build-in custom service program in the satellite server, the data can be securely "pushed" back through a firewall to the STAR Lab data warehouse using Open Database Connectivity (ODBC). This is more expensive but more secure solution to transmitting the data. Example:

- *Intersection detector event data:* Second-by-second event data are collected from all video sensors at the intersection at 196th Street and SR99 in the City of Lynnwood, WA. The data are stored in the STAR Lab satellite server and concurrently pushed through firewalls to the DRIVE Net Intersection Performance database.

4) Direct data archiving: The data can be collected directly from the data collection devices. The data can be sent directly and periodically to the data warehouse from the test site. Example:

- *Route travel time data:* Bluetooth-based travel time detectors developed by the STAR Lab can effectively collect route travel times by matching the unique Median

Access Control (MAC) address at various locations (Malinovskiy et al., 2011). This device is able to transfer data using General Packet Radio Service (GPRS) and Global System for Mobile Communications (GSM) communication protocols in real time. The data are sent directly back to the DRIVE Net Bluetooth travel time database every five minutes.

For the databases mentioned above, the schemas have been designed in advance to ensure data management and query efficiencies. The relational data model (Codd, 1970) is used in the design. All kinds of transportation data can be systematically stored in the DBMS and the relationships between the attributes (columns) can be easily maintained following the designed schema.

DRIVE Net Web Server

The core DRIVE Net system lies in the web server running Apache Tomcat 6.0 in the Windows Server 2008 Operating System (OS) environment. This server can render/disseminate the data and execute analytical algorithms depending on the role of users. Traffic engineers, researchers and travelers are three users groups expected to use DRIVE Net. For example, certain downloading functions are limited to certain user groups. As illustrated in Figure 6-1, DRIVE Net can be connected to multiple data servers using different data communication techniques. When necessary, another server can be added to the system as well. DRIVE Net servers will work jointly like terminals in the grid computing infrastructure.

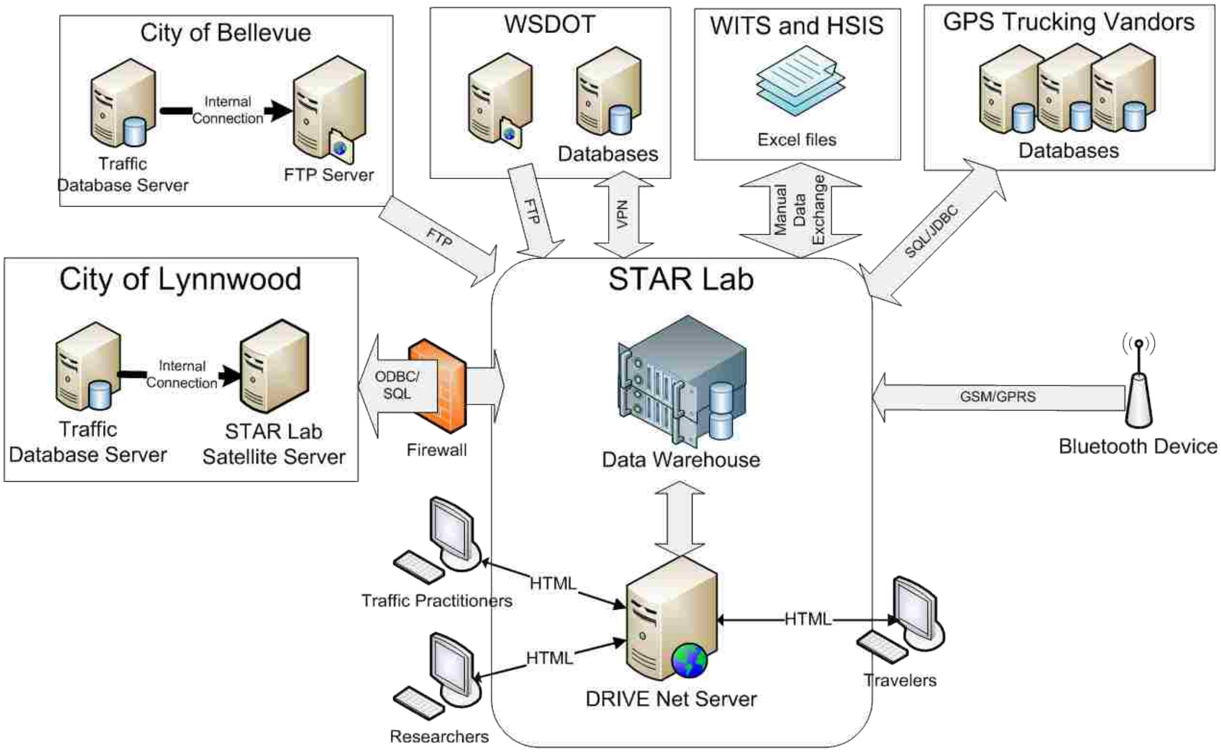


Figure 6-1DRIVE Net System Architecture

6.2.3 System Design

The DRIVE Net system is developed based on a multitier architecture model, commonly used in software engineering (Ran et al., 1999; Eckerson et al., 1995). The major merit of the multitier architecture is that the developers can modify or add a specific tier without rewriting the entire application. The model being used consists of a client-side presentation tier (client side web browser), a server-side data tier (data warehouse), and two server-side logic tiers (middleware and computational module). In addition to the traditional three-tier client-server model, an additional logic tier is added to handle data quality issues. The computational tier is used for data sharing control and algorithm execution. The middleware tier is designed to mitigate the burden in the computational tier. The burden is usually caused by excessive access to database, analytical algorithm calculations, and data quality control. The presentation tier is on the client side web browsers, and used for displaying interfaces and visualizing outputs, and receiving inputs from users. The overall system flow chart is shown in Figure 6-2:

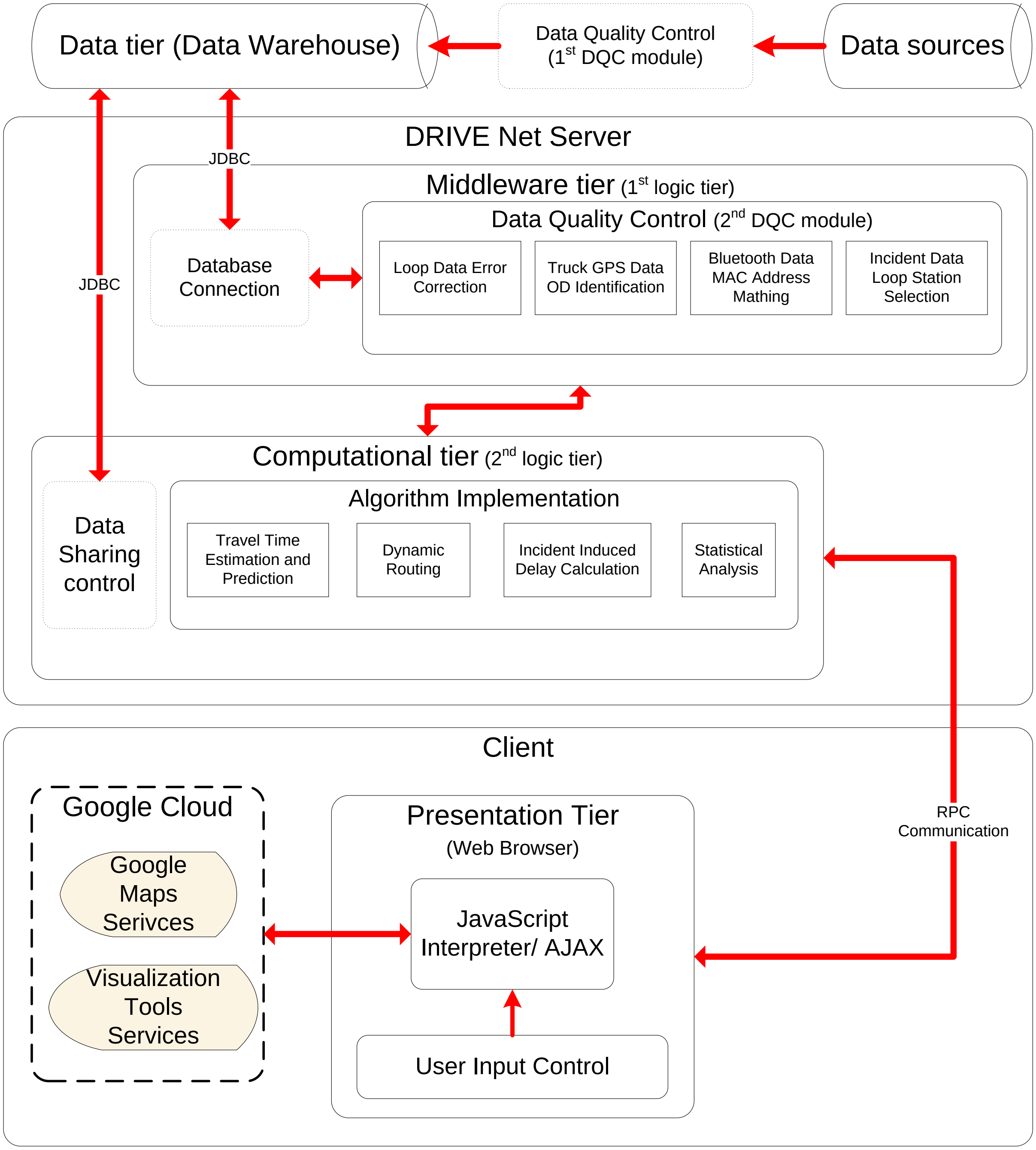

Data tier (Data Warehouse)
Data Quality Control (1st DQC module)
Data sources
DRIVE Net Server
JDBC
JDBC
Middleware tier (1st logic tier)
Data Quality Control (2nd DQC module)
Database Connection
Loop Data Error Correction
Truck GPS Data OD Identification
Bluetooth Data MAC Address Mathing
Incident Data Loop Station Selection
Computational tier (2nd logic tier)
Data Sharing control
Algorithm Implementation
Travel Time Estimation and Prediction
Dynamic Routing
Incident Induced Delay Calculation
Statistical Analysis
Client
RPC Communication
Google Cloud
Google Maps Serivces
Visualization Tools Services
Presentation Tier
(Web Browser)
JavaScript Interpreter/ AJAX
User Input Control

Figure 6-2 DRIVE Net System Flow Chart

Data Quality Control

Data quality is an issue that is widely recognized by transportation researchers and agencies. Developing an automatic and robust data quality control (DQC) procedure is beneficial to facilitate transportation-related research. To insure the quality data, DRIVE Net incorporates a two-step DQC mechanism handling data cleansing tasks, including error detection, removal and inconsistencies, etc. (Rahm and Do, 2000). The first step of data cleansing service happens at the stage of data retrieval from different data sources. For example, erroneous data are either flagged or removed. Examples of erroneous data include zero occupancy and negative volume in the loop detector data, and offset GPS data in the freight database. Further data cleansing (i.e. second step data cleansing) is handled in the DQC module in the middleware tier. In addition to error checking, DQC in the middleware tier mainly conducts preliminary data analysis and processing to reduce the computational burden in the computational tier. For example, the advance loop detectors at Bellevue's intersections are wired together, resulting in undercount problems. A probability-based nonlinear model developed by (Wu et al., 2010) is incorporated into the second DQC module to correct the undercounted volume. Freeway ILD data suffers from both misdetections and erroneous occupancy issues due to incorrect sensitivity level settings in loop cards (Cheevarunothai et al., 2006). A software-based error detection and correction algorithm (Wang et al., 2009) is also implemented in the middleware tier. Another example is the Origin-Destination (OD) identification algorithm developed by (Ma et al., 2011), which incorporated and extracted individual truck OD information for freight performance measurement. Similarly, the raw Bluetooth MAC addresses collected by the Bluetooth detectors are sent back to DRIVE Net. The redundant data were screened at the first DQC module and the travel time calculations are also undertaken in the 2nd DQC module of the middleware tier.

Middleware Tier

Middleware is a computer program independently running in the server. As mentioned, the purpose of building a middleware tier is to leverage computational power, manage resources between the server (data and two logic tiers) and the client (presentation tier). In addition to the DQC module mentioned earlier, the data connection module is also developed in the middleware tier. In fact, this module is a program interface to connect with multiple databases using Java Database Connectivity (JDBC) API, allowing the middleware tier to query and receive the results from the data warehouse for further process.

Computational Tier

The computational tier in the DRIVE Net server handles complex algorithm implementation after DQC is complete. In addition, this tier assists in archiving raw data and data sharing service control. The Asynchronous JavaScript and XML (AJAX) (Garrett, 2010) technology is implemented to reduce the data transfer between the server and the browser and minimize interference to the display and ongoing activities on the existing page. This design reduces the server's response time and enhances the system performance for displaying dynamic and interactive web pages (Garrett, 2010).

Multiple algorithms implemented in DRIVE Net use this AJAX technology. These algorithms include a iterative time-dependent A* algorithm performing the shortest travel time routing (Wu et al., 2011), statistical metrics generation for freight performance measures (Ma et al., 2011) and incident induced delay calculation using deterministic queue theory and time series techniques (Yu et al., 2011).

Presentation Tier

The primary functionality of the client side is to provide an interactive Graphic User Interface (GUI). As shown in Figure 6-2, the users' inputs are sent to the computational tier. The computed results are then sent back to the web browser through the Remote Procedure Call (RPC). The final results are visualized through Google Map API (Google, 2010) and Visualization API (Google, 2010), two major third party components supported by the Google cloud. The Google Maps API allows developers to visualize the results on Google Maps through Google Maps services. The Google Visualization API allows users to interact with the data visualized in the statistical charts, such as histograms and pie charts, through visualization tolls services.

Implementation

A combination of the Google Web Toolkit (GWT) (Google, 2010) and Eclipse (Eclipse, 2010), an open source Integrated development environment (IDE), creates a strong development environment for DRIVE Net. GWT contains Java API libraries, allowing developers to code web applications in Java language and then compile the source code into JavaScript. In this case, development cost and time are significantly reduced compared with traditional web development methods, such as JavaScript and/plus PHP. In addition, debugging in GWT makes traditional JavaScript web development much convenient. A developer is able to access existing widget templates in the GWT library to design web interfaces, and a Java to JavaScript compiler translates and optimizes Java code into JavaScript. The prototype DRIVE Net system can be accessed online at http://www.uwdrive.net/. The web interface of DRIVE Net (Version 2) is shown in Figure

6-3. All computational functions are located on the left side of the panel, including total eight modules programmed on an objected-oriented basis. Hence, all the classes can be “recycled” and “reused” for future development.

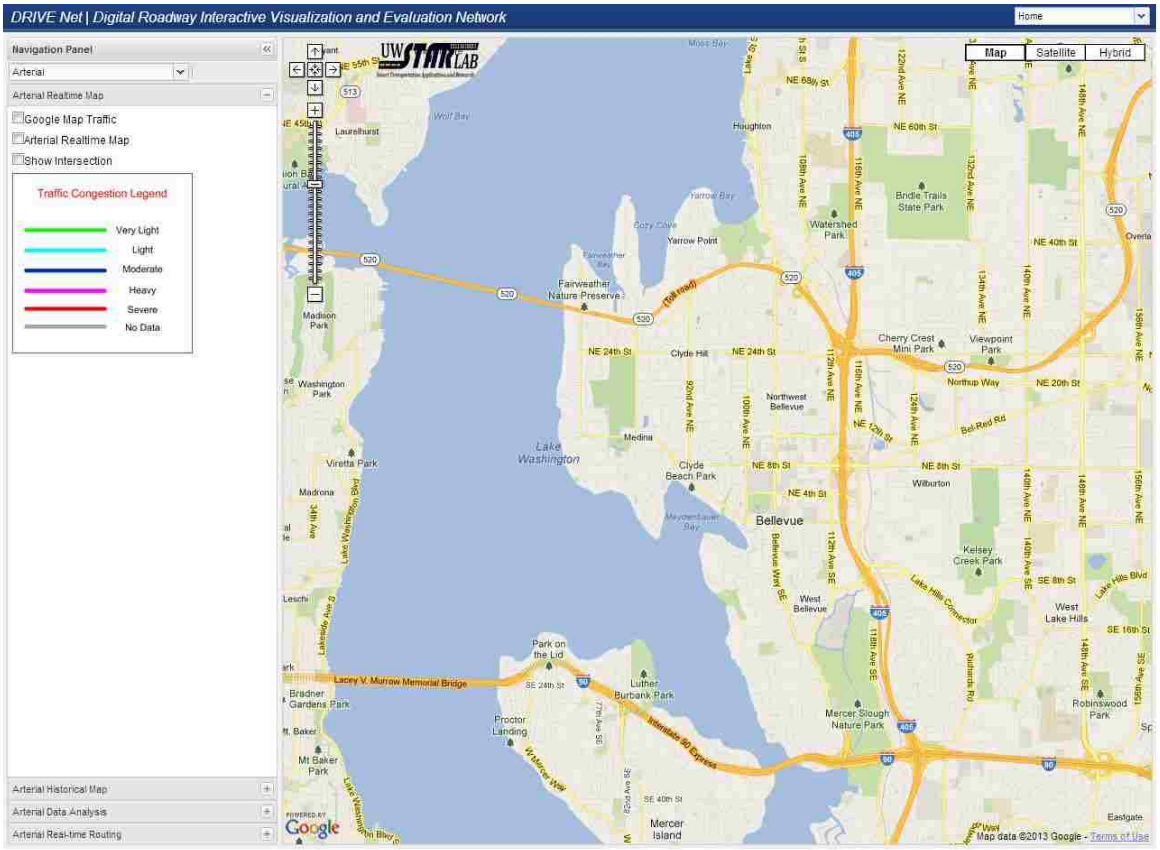


Figure 6-3 DRIVE Net Interface (Version 2)

6.3 TransitNet: An E-Science Transportation Platform for Transit Performance Measures

Although DRIVE Net system (version 2) established an interoperable transportation data framework to enhance connections of heterogeneous data sources, it still leaves a critical issue on how to integrate with geospatial data in an efficient and effective manner. The existing transportation GIS frameworks are not satisfying due to the following two reasons: (1) lack an explicit geospatial data model to integrate both transportation information and roadway geometric features; (2) slow performance for online transportation GIS applications due to redundancy. In this section, a light-weight transportation GIS data model is proposed, and it is particularly designed for web-based applications by utilizing several open-source software tools. As an improved version of DRIVE Net (version 2), TransitNet incorporates this simplified transportation GIS data model into its development, and can better represent the large-scale transit network for transit performance measures.

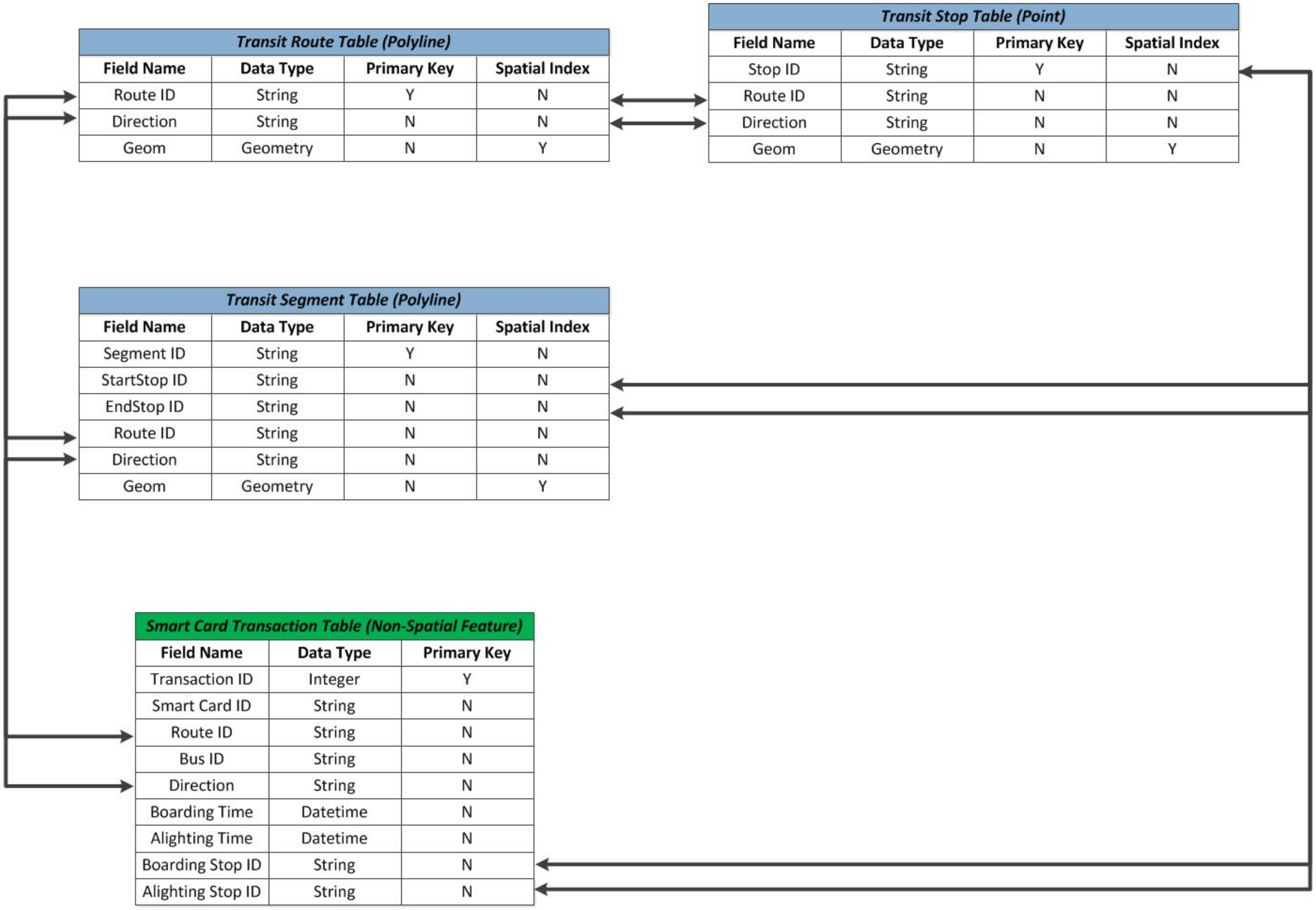
6.3.1 A Simplified Transportation GIS Data Model

A transit network is primarily composed of transit routes and transit stops. Transit stops are located along each particular transit route, and certain transit stops are served as transfer points to connect different transit routes. Traditional GIS applications treat transit stops and transit routes as points and lines respectively, and store them as shapefiles (ESRI, 1998). A shapefile contains the geospatial attributes such as points, polylines, and polygons, and additional non-spatial features can be also associated with the shapefile, for example, the population of an area and the length of a river. Similar concepts can be applied to represent a transit network, where transit stops and transit routes can be archived as shapefiles. However, the non-spatial attributes of a transit network cannot be easily processed in shapefiles. This is because the non-spatial attributes of either a transit stop or a transit route vary by time. A typical example is the ridership changes for a particular transit stop. The number of boarding passengers is not a constant, and thus it is difficult to use a static file to store this dynamic information.

The emergence of geospatial database techniques can alleviate the burden of file-based geospatial data management and analysis. Similar as the traditional Relational DataBase Management System (RDBMS), Geospatial databases can optimize the geospatial data management and analysis by using Structured Query Language (SQL) techniques and spatial indices. In addition, geospatial databases enable a variety of geo-processing operations that traditional relational non-spatial databases cannot be complete, for example, whether two polylines intersect with each other, or whether points fall within a spatial area of interest. However, in reality, most transportation agencies utilize non-spatial relational databases to store traffic-related information such as transit GPS data and transit smart card transactions data. This creates a critical issue: how to best represent and manage the dynamic transportation data in a context of hybrid spatial and non-spatial databases. Especially when more and more location-aware transportation data are available for advancing Big Data initiative, this issue becomes more pressing.

Although a wealth of commercialized software packages has provided various solutions to tackle this issue, usability and accessibility are far from satisfaction. These packages only offer limited functionalities to a certain group of experienced GIS users, and cannot satisfy majority of transportation professions for specific and customized purposes. Moreover, when integrated with the Internet, long response time can be seen in most of the commercialized software packages for online GIS applications. This is probably because: (1) there are too many unnecessary modules during the loading process; (2) the transportation GIS model in these packages is not well designed. To address these issues, a simplified and flexible transportation GIS data model is proposed for both transportation data management and visualization purposes.

The first challenge is how to represent the fundamental geometry features in geospatial databases in an efficient and effective fashion. Each geometric record (e.g. polygon, polyline and point) in the geospatial database are recoded in a manner of Well-known binary (WKB). WKB encoding method is defined by the Open Geospatial Consortium (OGC) (ISO and IEC, 2011), and uses the binary content to encode vector geometry. Both the coordinate (e.g. latitude and longitude) and projection information are include in a hexadecimal string, and can be parsed as a sequence of latitude and longitude pairs to visualize on any mapping system. For example, the stop A is a point geometry with the latitude and longitude pair as (116.564, 40.009), and it can be represented as a string of "0101000020E61000003C3AA9A018245D4016EDDB4C21014440" in the geospatial database. In this study, a transit segment is defined as the line between two adjacent transit stops, and considered as a fundamental element to conduct the network-level and route-level performance measures. To accomplish this task, each transit route is segmented by several transit stops as a sequence of short links, and each link is bounded with the adjacent transit stops. The simplified transit GIS data model can be expressed as in figure 6-4:



Transit Route Table (Polyline)			
Field Name	**Data Type**	**Primary Key**	**Spatial Index**
Route ID	String	Y	N
Direction	String	N	N
Geom	Geometry	N	Y

Transit Stop Table (Point)			
Field Name	**Data Type**	**Primary Key**	**Spatial Index**
Stop ID	String	Y	N
Route ID	String	N	N
Direction	String	N	N
Geom	Geometry	N	Y

Transit Segment Table (Polyline)			
Field Name	**Data Type**	**Primary Key**	**Spatial Index**
Segment ID	String	Y	N
StartStop ID	String	N	N
EndStop ID	String	N	N
Route ID	String	N	N
Direction	String	N	N
Geom	Geometry	N	Y

Smart Card Transaction Table (Non-Spatial Feature)		
Field Name	**Data Type**	**Primary Key**
Transaction ID	Integer	Y
Smart Card ID	String	N
Route ID	String	N
Bus ID	String	N
Direction	String	N
Boarding Time	Datetime	N
Alighting Time	Datetime	N
Boarding Stop ID	String	N
Alighting Stop ID	String	N

Figure 6-4 A Simplified Transit GIS Data Model

As shown in Figure 6-4, the field “Geom” defines the geometry attribute of each table, and it is stored as a form of WTK in geospatial databases. Foreign keys are used to link different geospatial tables for cross-reference. For example, both transit route table and transit stop table share the common fields of route ID and direction. As a derived table using the relationship between a transit route and transit stops, transit segment table can act a mediator to represent the entire transit network for segment-based travel speed visualization. All the above transit spatial data are imported and managed into a geospatial database. Meanwhile, the origin and destination information for each smart card transaction is updated in a non-spatial database by applying the aforementioned data mining approaches. To bridge the spatial data and non-spatial data in a loose-coupled manner, common fields are also utilized to merge heterogeneous datasets. This integration method behaves like the traditional database “join” query by combining the common fields from different types of databases. For example, both the route ID and direction fields from the smart card transaction table can associate with the same fields from the transit route table. Similarly, the boarding stop ID and the alighting stop ID of each smart card transaction should correspond to the startstop ID and the endstopID in the transit segment table respectively. To reduce the computational expense for searching and matching the common attributes, Hash table data structure can be implemented by creating a hash function for indexing between both tables.

In summary, this simplified transit GIS data model can efficiently manage the relationship between spatial data (transit network elements) and transportation data (transit smart card data), and it only requires minimum information to couple with the two types of datasets. Therefore, it is particularly suitable to handle a large quantity of transportation data such as Beijing transit smart card data without incurring too much redundancy. Both the transit network geospatial data and smart card data can be independently processed in this context.

6.3.2 Migrate DRIVE Net to TransitNet: An Improved System Design

The proposed transit GIS data model can be incorporated in the design of TransitNet system. In addition, several new features are introduced into TransitNet as well. Figure 6-5 demonstrates the improved system design of TransitNet:

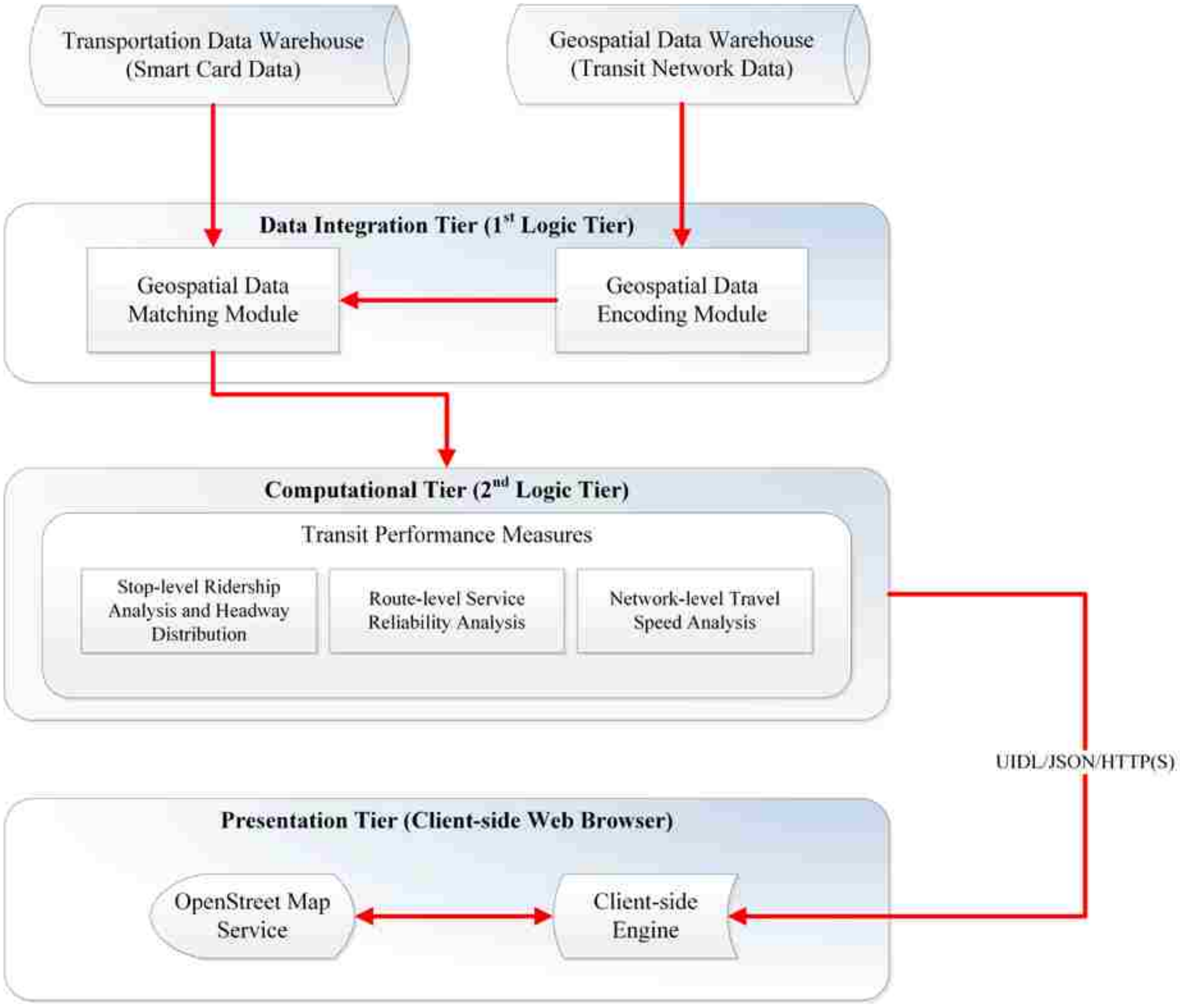

Transportation Data Warehouse (Smart Card Data)
Geospatial Data Warehouse (Transit Network Data)
Data Integration Tier (1st Logic Tier)
Geospatial Data Matching Module
Geospatial Data Encoding Module
Computational Tier (2nd Logic Tier)
Transit Performance Measures
Stop-level Ridership Analysis and Headway Distribution
Route-level Service Reliability Analysis
Network-level Travel Speed Analysis
UIDL/JSON/HTTP(S)
Presentation Tier (Client-side Web Browser)
OpenStreet Map Service
Client-side Engine

Figure 6-5 System Design of TransitNet

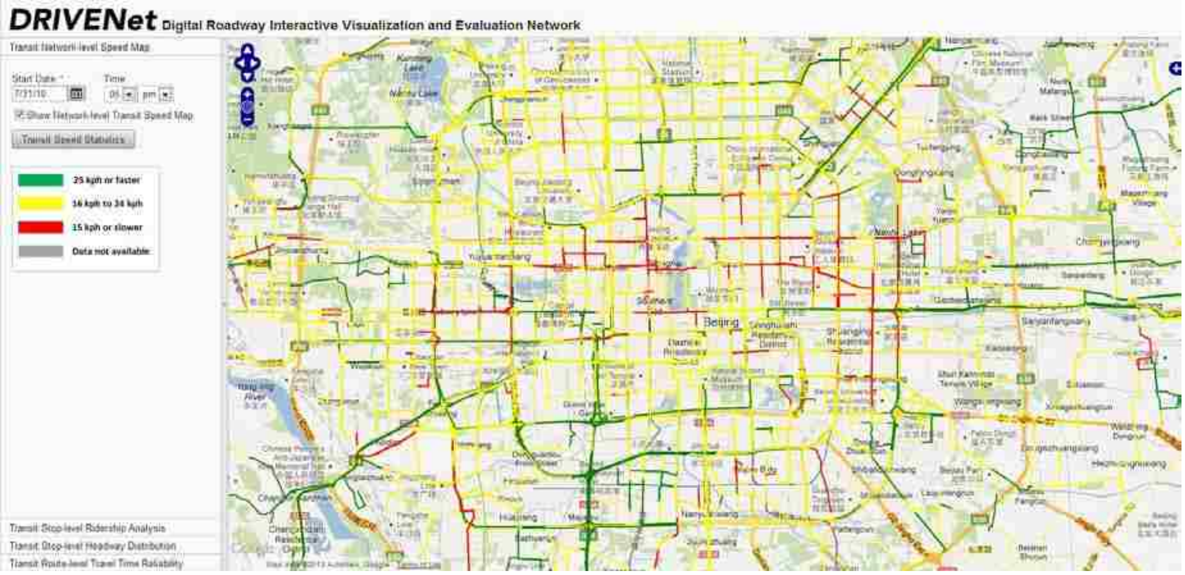
Compared with the system design of DRIVE Net system, the data integration tier is an additional functionality. This tier is equivalent to the application of the proposed transit GIS data model. Transit network spatial data (route, stop and segment) are digested into the geospatial data encoding module for WTK format conversions, and then associate with the common fields from the no-spatial smart card database. The matched geospatial smart card data are extracted to conduct further transit performance metric calculation at different scales. The calculated statistics is finally sent to the client-side web browsers to display, such as ridership temporal histogram at a particular stop, color map of network-wide transit speed, etc. The client-side engine is responsible to parse the geometry feature of each statistical result (e.g. a congested transit segment) as a sequence of latitude and longitude pairs, and presented in an OpenStreetMap interface for visualization purposes. The entire system is implemented using several open-source software packages: PostgreSQL (The PostgreSQL Global Development Group, 2013) and PostGIS (Refractions Research, 2013) are adopted to store the transit spatial information. Vaddin (Vaadin Ltd., 2013) is used to construct the interactive graphical user interface in Java.

6.3.3 Key Components in TransitNet

As mentioned in section 6.1, a hierarchical transit performance measurement framework is proposed called TransitNet. TransitNet can be accessed at www.uwdrive.net/TransitNet. In this section, the four key components of TransitNet are presented as below:

Transit Network-level Speed Map

The transit travel speed can be calculated using the identified passenger OD pairs. Figure 6-6 demonstrates the Beijing transit network traffic condition from 4:30PM to 5:00PM on July, 30, 2010 (weekday).



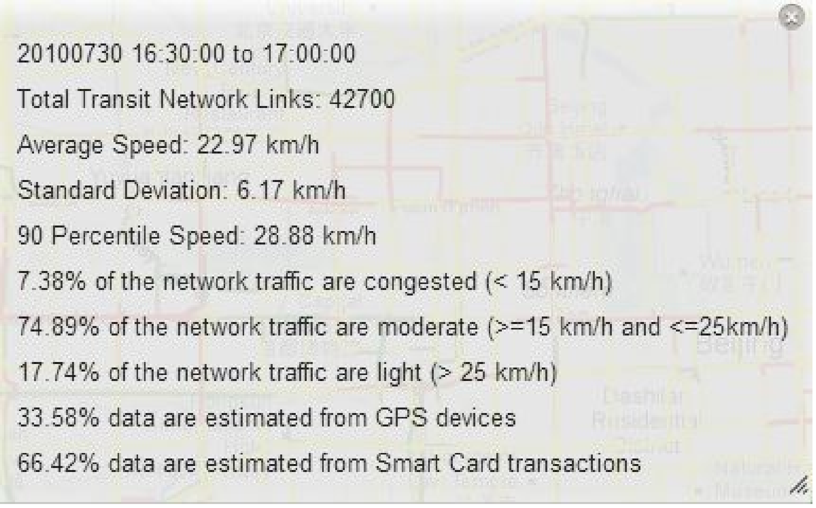
DRIVENet Digital Roadway Interactive Visualization and Evaluation Network
Transit Network-level Speed Map
☑ Show Network-level Transit Speed Map
Transit Speed Statistics
🟩 25 kph or faster
🟨 16 kph to 24 kph
🟥 15 kph or slower
▨ Data not available
Transit Stop-level Ridership Analysis
Transit Stop-level Headway Distribution
Transit Route-level Travel Time Reliability
Kunming Lake
Beijing

Start Date: 7/31/10

Time: 05 pm

Figure 6-6 Beijing Transit Network Speed Map

Additional analysis functionality is provided to view the network-wide transit travel speed statistics. By clicking transit speed statistics button, a window (as showed in Figure 6-7) is popped up to illustrate the detailed transit speed information for the entire network, such as average speed, deviation, 90th percentile speed, percentages of uncongested/congested transit segments, and the composition of data sources for the transit speed calculation (GPS and smart card). The colored speed map and statistical analysis can be used as effective tools for transit agencies to identify congested areas, and then improve their public transit services accordingly such as opening express lanes for public transit or shortening headways, etc.



20100730 16:30:00 to 17:00:00

Total Transit Network Links: 42700

Average Speed: 22.97 km/h

Standard Deviation: 6.17 km/h

90 Percentile Speed: 28.88 km/h

7.38% of the network traffic are congested (< 15 km/h)

74.89% of the network traffic are moderate (>=15 km/h and <=25km/h)

17.74% of the network traffic are light (> 25 km/h)

33.58% data are estimated from GPS devices

66.42% data are estimated from Smart Card transactions

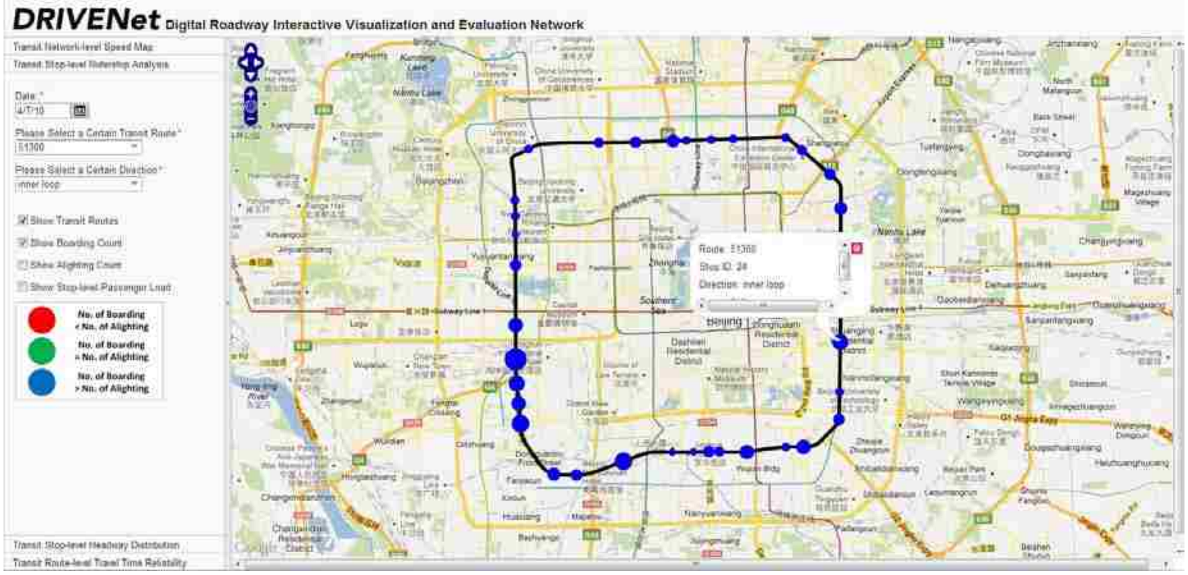
Figure 6-7 Network-wide Transit Speed Statistics Windows

As presented in Figure 6-6 and Figure 6-7, the network-wide transit average speed is 22.97 km/h, and its standard deviation is around 6.17 km/h. More than 80% transit links are highly congested or moderately congested, where the transit speed is under 25 km/h. severe congestions can be found in the central district of Beijing through the transit speed spatial distribution. 66% of speed calculations are from smart card transactions, and 33% of calculations are from GPS devices. This is a reasonable result for the afternoon peak hour traffic in Beijing. Most roadways suffer from congestions and thereby low speed can be observed as those red lines in the map.

Transit Stop-level Ridership Analysis

Transit ridership analysis is of critical significance for transit agencies. Transit agencies strive to retain the existing transit riders, and also attract more potential transit riders. Presenting the spatiotemporal transit ridership through a map-based system will be beneficial for transit agencies to conduct the before-and-after analysis and understand passenger demand changes. An example is shown in figure 6-8.

Route 51300 is a flat-fare based loop route with a total of 34 transit stops. Both the passenger boarding and alighting stops can be inferred by the aforementioned smart card OD estimation approaches. To visualize these results, three stop-level analytical tools are provided: the number of boarding passengers, the number of alighting passengers and the passenger load profile.


DRIVENet Digital Roadway Interactive Visualization and Evaluation Network
Transit Network-level Speed Map
Transit Stop-level Ridership Analysis
Transit Stop-level Headway Distribution
Transit Route-level Travel Time Reliability
Beijing

Date:* 4/7/10

Please Select a Certain Transit Route:* 51300

Please Select a Certain Direction:* inner loop

☑ Show Transit Routes

☑ Show Boarding Count

☐ Show Alighting Count

☐ Show Stop-level Passenger Load

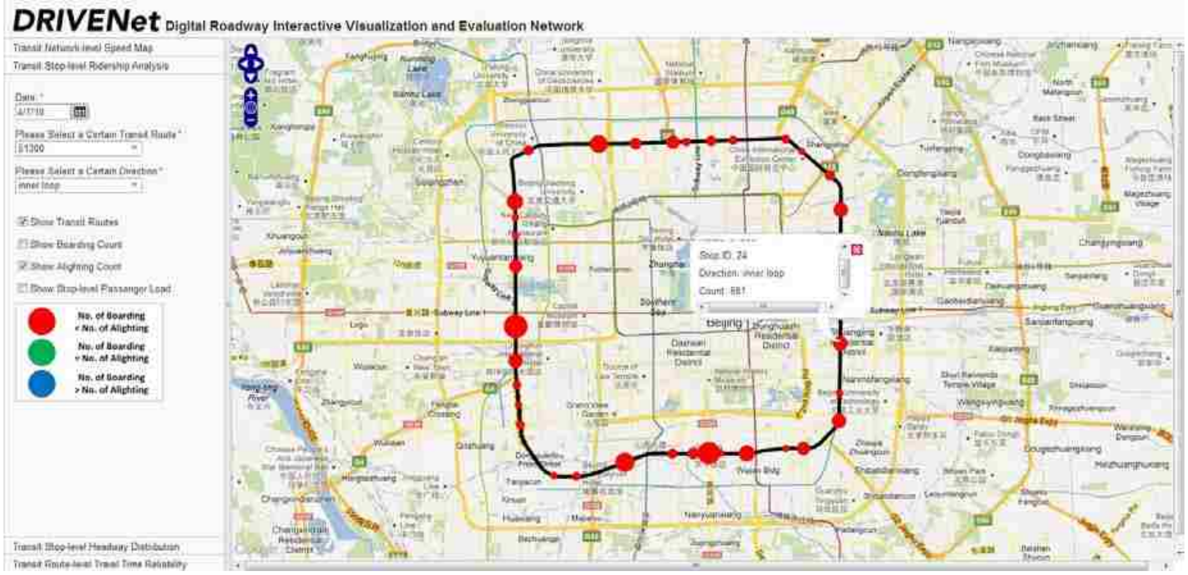
🔴 No. of Boarding < No. of Alighting

🟢 No. of Boarding = No. of Alighting

🔵 No. of Boarding > No. of Alighting

Route: 51300
Stop ID: 24
Direction: inner loop

Figure 6-8 Number of Boarding Passengers Distribution



DRIVENet Digital Roadway Interactive Visualization and Evaluation Network

Transit Network-level Speed Map

Transit Stop-level Ridership Analysis

Date: * 4/7/10

Please Select a Certain Transit Route * 51300

Please Select a Certain Direction * inner loop

☑ Show Transit Routes

☐ Show Boarding Count

☑ Show Alighting Count

☐ Show Stop-level Passenger Load

🔴	No. of Boarding < No. of Alighting
🟢	No. of Boarding = No. of Alighting
🔵	No. of Boarding > No. of Alighting

Transit Stop-level Headway Distribution

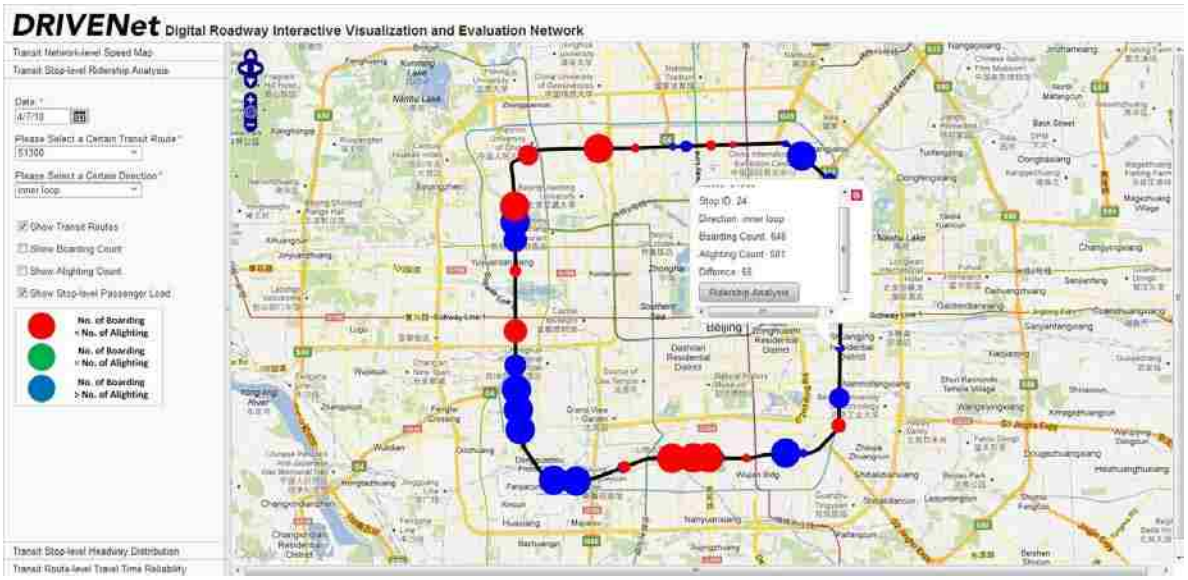
Transit Route-level Travel Time Reliability

Stop ID: 24

Direction: inner loop

Count: 581

Figure 6-9 Number of Alighting Passengers Distribution



DRIVENet Digital Roadway Interactive Visualization and Evaluation Network

Transit Network-level Speed Map

Transit Stop-level Ridership Analysis

Date: * 4/7/10

Please Select a Certain Transit Route * 51300

Please Select a Certain Direction * inner loop

☑ Show Transit Routes

☐ Show Boarding Count

☐ Show Alighting Count

☑ Show Stop-level Passenger Load

🔴	No. of Boarding < No. of Alighting
🟢	No. of Boarding = No. of Alighting
🔵	No. of Boarding > No. of Alighting

Transit Stop-level Headway Distribution

Transit Route-level Travel Time Reliability

Stop ID: 24

Direction: inner loop

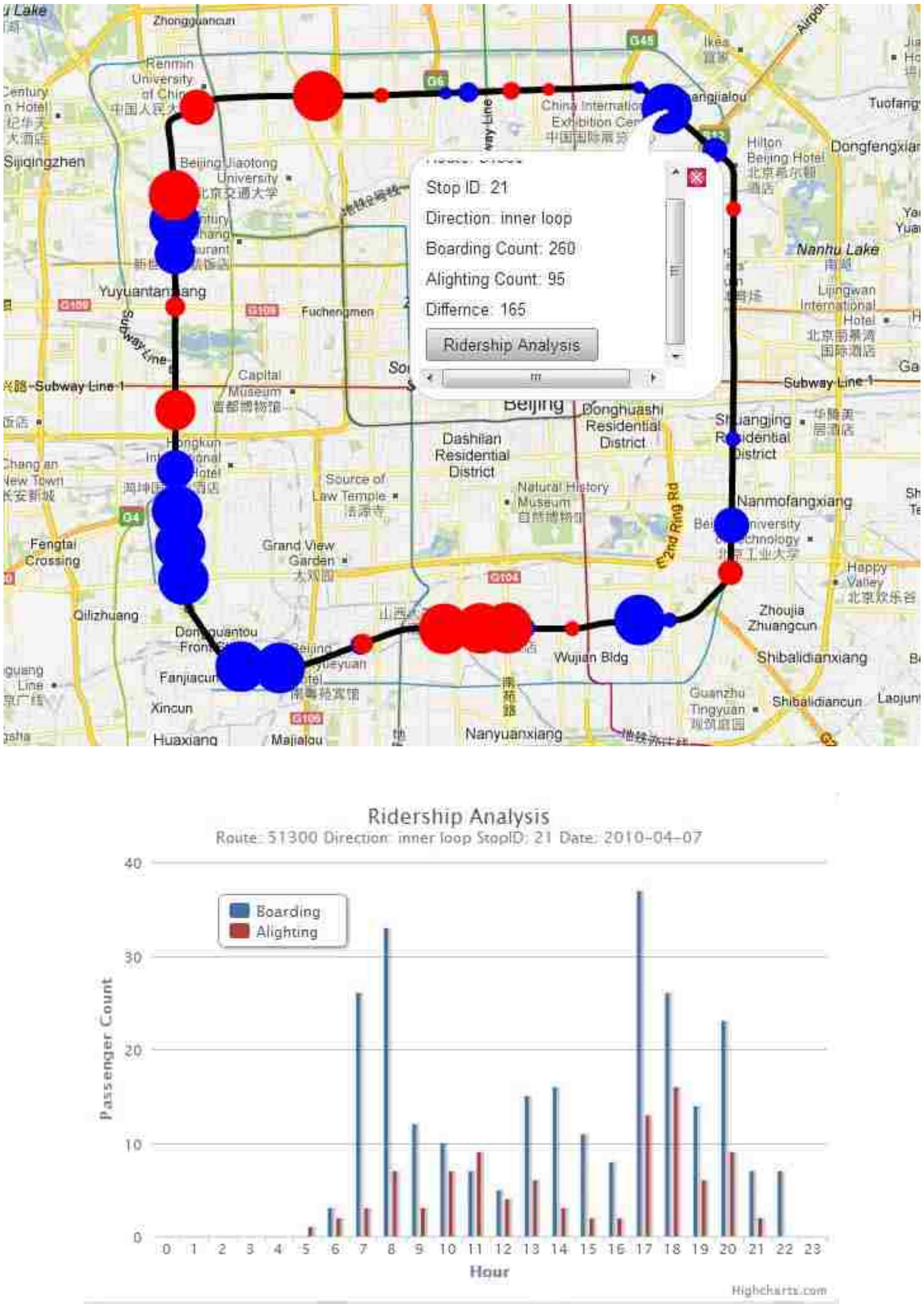
Boarding Count: 646

Alighting Count: 581

Diffrence: 65

Figure 6-10 Passenger Load Distribution

Number of boarding passengers and number of alighting passengers are colored in blue and red respectively shown in Figure 6-8 to Figure 6-10. The radius of each stop represents the magnitude of passenger counts. The larger the radius is, the more passengers board or alight at each stop. Similarly, the passenger load at each stop is defined as the difference between the boarding passenger counts and the alighting passenger counts. If the total passenger boarding activities are more than the total passenger alighting activities at a certain stop, the level of congestion inside a transit vehicle at this stop will increase. Consequently, the comfort level of passengers may degrade, and could affect transit riders' mode choices in the future. The passenger load spatial distribution offers an intuitive method for transit agencies to understand stop-level passenger demands. In addition, TransitNet can also generate a fine-grained histogram chart for the temporal distribution of ridership at each stop. As shown in Figure 6-11, Stop ID 21 of route 51300 is selected as a transit stop with high passenger demands. To further investigate how the ridership varies by time. An interactive temporal histogram can be generated by clicking the button of "ridership analysis". Compared with the passengers who got off the buses, more passengers got on the buses from the morning peak hour (7:00 AM to 9:00 AM) and afternoon peak hour (5:00 PM to 7:00 PM). This may lead to a congested environment inside the buses, and further prevents passengers at upstream stops from boarding. Transit agencies can resort to this temporal tool to improve transit service availability by adding more buses during rush hours.



Stop ID: 21

Direction: inner loop

Boarding Count: 260

Alighting Count: 95

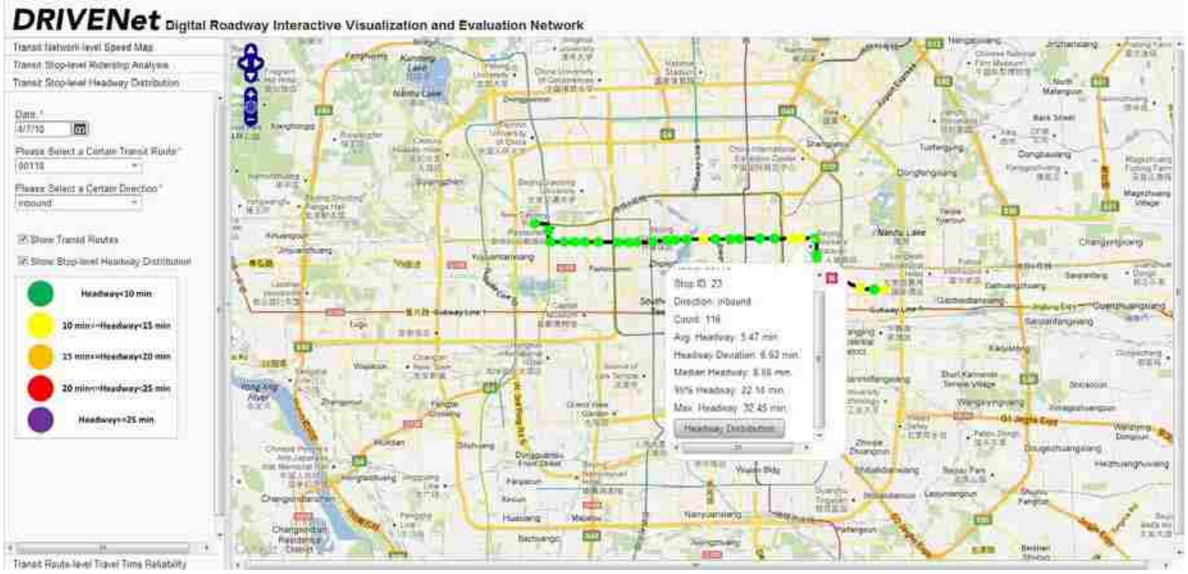
Differnce: 165

Ridership Analysis

Figure 6-11 Stop-level Ridership Analysis for Route 51300

Transit Stop-level Headway Distribution

Headway is another critical performance measure to quantify the transit service quality from transit riders' perspectives, and it can be considered as the maximum waiting time for each transit rider. Due to the traffic signal delay and dwell time at each transit stop, the headway for each stop may vary, and thus it is necessary and useful for transit agencies to visualize the statistics of stop-level headway for decision making. Figure 6-12 presents an example of route 118 on April, 7, 2008. A total 116 of bus runs can be estimated using the smart card transaction data, and the average headway and headway deviation for stop 23 can be further derived as 9.47 minutes and 6.63 minutes respectively. The mean headway is used to color each transit stop. In this case, transit agencies can easily identify those transit stops with the relatively long headways, and make corresponding schedule adjustments. A frequency histogram is also generated to depict the distribution of headway (Figure 6-13). For stop 23 in the route 118, most time intervals between two consecutive buses are small (around 8 minutes). Although there are several outliers (e.g. 31 minutes and 32 minutes), most of these long headways occurred during rush hours, and thus traffic congestion is the contributing factor to impact headways.



DRIVENet Digital Roadway Interactive Visualization and Evaluation Network
Transit Network-level Speed Map
Transit Stop-level Ridership Analysis
Transit Stop-level Headway Distribution
Transit Route-level Travel Time Reliability
Headway Distribution

Date: * 4/7/10

Please Select a Certain Transit Route * 00118

Please Select a Certain Direction * inbound

☑ Show Transit Routes

☑ Show Stop-level Headway Distribution

🟩 Headway<10 min

🟨 10 min<=Headway<15 min

🟧 15 min<=Headway<20 min

🟥 20 min<=Headway<25 min

🟪 Headway>=25 min

Stop ID: 23
Direction: inbound
Count: 116
Avg. Headway: 5.47 min
Headway Deviation: 6.63 min
Median Headway: 8.55 min
95% Headway: 22.14 min
Max. Headway: 32.45 min

Figure 6-12 Transit Stop-level Headway Spatial Distribution

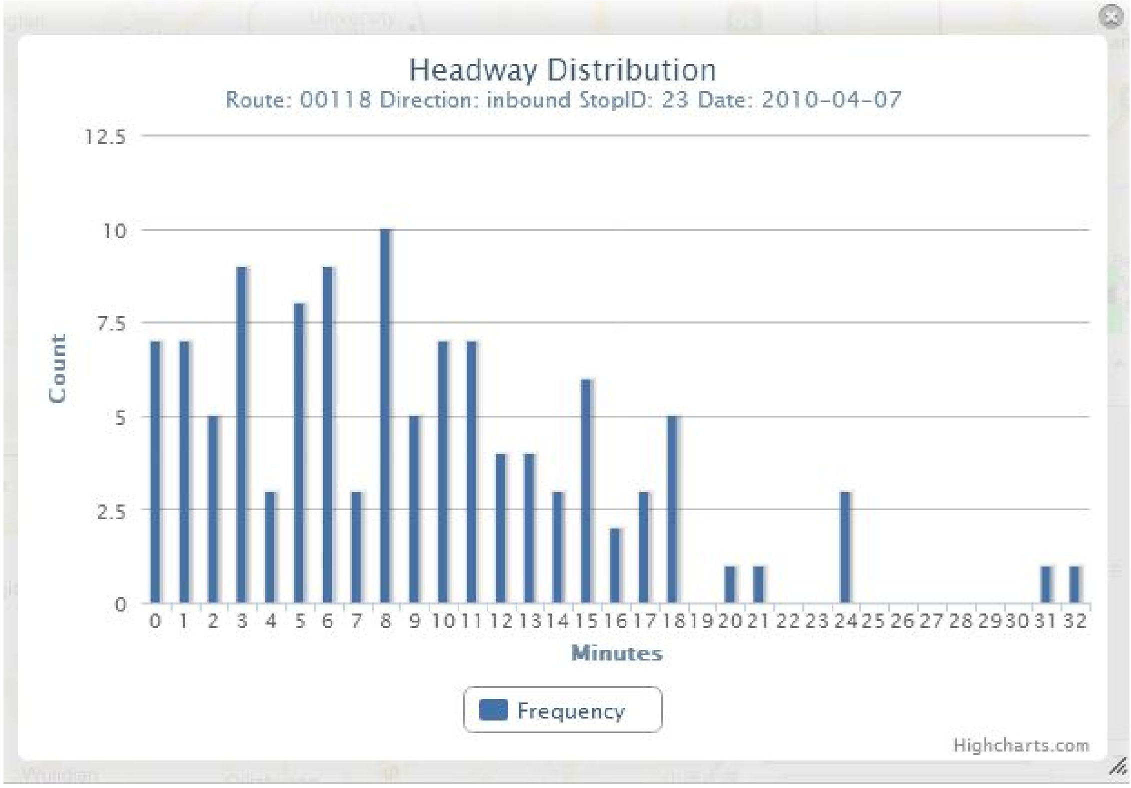
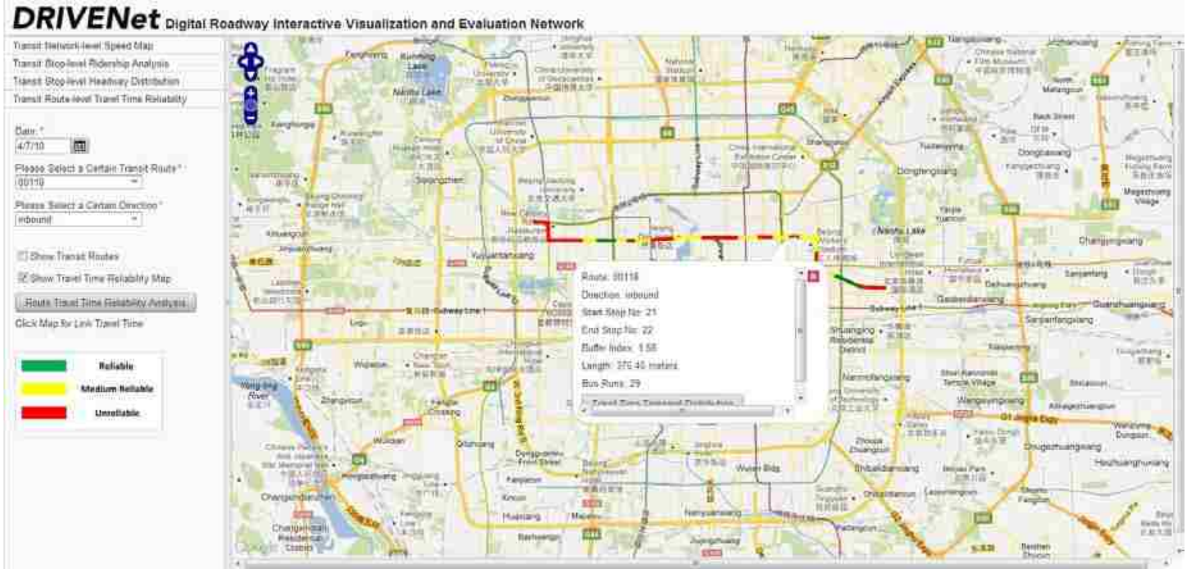


Figure 6-13 Transit Headway Histogram

Transit Route-level Travel Time Reliability

Another feature of TransitNet is to visualize the transit travel time reliability on a colored map (Figure 6-14). Buffer time index is used to measure the travel time reliability. The smaller the buffer time index is, the more reliable a transit route is.


DRIVENet Digital Roadway Interactive Visualization and Evaluation Network
Transit Network-level Speed Map
Transit Stop-level Ridership Analysis
Transit Stop-level Headway Distribution
Transit Route-level Travel Time Reliability
Route Travel Time Reliability Analysis
Click Map for Link Travel Time
Reliable
Medium Reliable
Unreliable
Travel Time Temporal Distribution

Date: *
4/7/10

Please Select a Certain Transit Route *
00118

Please Select a Certain Direction *
inbound

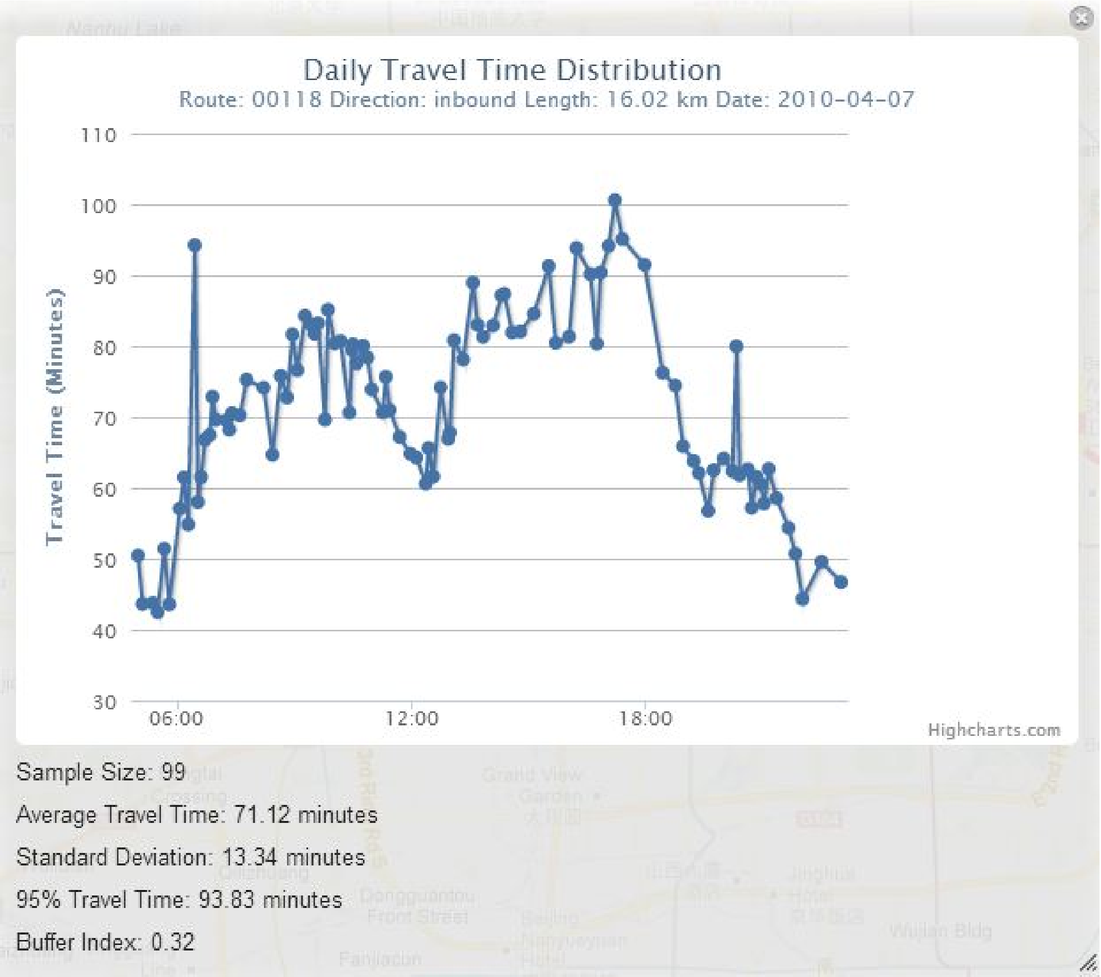
☐ Show Transit Routes
☑ Show Travel Time Reliability Map

🟩 Reliable
🟨 Medium Reliable
🟥 Unreliable

Route: 00118
Direction: inbound
Start Stop No: 21
End Stop No: 22
Buffer Index: 1.55
Length: 376.45 meters
Bus Runs: 29

Figure 6-14 Route-level Transit Travel Time Reliability Spatial Distribution

To visually analyze the transit travel time change, the travel time of each transit route/link can be also represented as Figure 6-15. For route 118 on April 7, 2010, the buffer time index for the entire route is 0.32, which implies that transit rider should allow 22.71 minutes to ensure on-time arrival 95 percent of the time.



Sample Size: 99

Average Travel Time: 71.12 minutes

Standard Deviation: 13.34 minutes

95% Travel Time: 93.83 minutes

Buffer Index: 0.32

Figure 6-15 Transit Route Travel Time Trend Analysis

Chapter 7 Conclusions and Future Research

7.1 Conclusions

Public transit is considered as an effective countermeasure to alleviate congestion and to reduce emissions and energy consumption. Therefore, improving public transit quality of service and ultimately attracting more ridership is of critical significance for transit agencies, and it requires methodological methods to better quantify transit service for operation planning and system optimization. However, historically, monitoring transit system performance has not been a straightforward task due to a scarcity of data. Recent passive transit data collection techniques (e.g. automatic fare collection system and automatic vehicle location system) have shifted public transit system to a data-rich platform, and enabled more opportunities and challenges to conduct data-driven research.

This study utilized Beijing smart card data to demonstrate the feasibility of establishing a web-based E-Science system for transit performance measures. Different from most entry-only AFC systems in other countries, Beijing's AFC system does not record boarding location information when passengers get on the buses and swipe their smart cards. Also, there are only a limited number of buses equipped with GPS devices. This creates challenges for passenger OD estimation.

The first task is to infer passenger origin information using smart card and GPS data. For those buses with GPS devices, stop-level arrival time can be efficiently estimated by integrating GPS records and stop location data. Each smart card transaction record is then associated with the inferred bus arrival time to identify the missing boarding stop. For those buses without GPS devices, a Bayesian decision tree algorithm is proposed to infer passenger origin stop, This algorithm is based on Bayesian conditional probability theory, and the probability density function of the segment-based travel speed is used to measure the randomness of passenger boarding stops, where its mean and variance are not sensitive to the algorithm accuracy and thereby not dependent on other data sources. Moreover, we can use the time invariance the of Markov chain model to further reduce the computational complexity of the algorithm to linear from exponential. The optimized algorithm has proven its accuracy and efficiency using the SC transaction data from two routes.

Then, to better understand each transit rider’s spatial and temporal travel patterns and regularity, a series of data mining procedures through multiday smart card transactions was developed. With a better understanding of the travel patterns and regularity (the “magnitude” level of travel pattern) of transit riders, transit authorities can evaluate the current transit services to adjust marketing strategies, keep loyal customers and improve transit performance. However, it is fairly challenging to identify travel patterns for each individual transit riders in a large dataset. Therefore, this study proposes an efficient and effective data-mining procedure that models the travel patterns of transit riders. Transit riders’ trip chains are identified based on the temporal and spatial characteristics of smart card transaction data. Based on the identified trip chains, the Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is used to detect each transit rider’s historical travel patterns. The K-Means++ clustering algorithm and the rough-set theory are jointly applied to clustering and classifying the travel pattern regularities. The rough-set-based algorithm is compared with other prevailing classification algorithms. The results demonstrate both the efficiency and accuracy of the proposed algorithm, and illustrate the feasibility of applying to very large datasets.

On the basis of the mined travel patterns, individual level passenger destination stop can be further estimated considering both the temporal and spatial relationship: Spatiotemporal transfer activities between different transit routes are firstly utilized to infer the possible alighting stops for passengers, and then, each individual's daily trip chain information (e.g. commuting behavior) is also taken into account to infer the missing alighting stops. Furthermore, because the historical travel patterns for each passenger are successfully extracted through weekly smart card transaction data, behavioral information can be also incorporated into passenger destination estimation to gain more insights. The proposed approach is tested using the distance-based fare smart card data from two routes, and more than 70% of the alighting stops can be accurately estimated within one-stop-error range.

To convey transit-related information in a timely and understandable manner, a web-based E-Science platform for visualizing, modeling, and analyzing transportation data was developed in this study. This platform ties transportation data with geospatial data using an efficient and effective GIS engine, and demonstrates several transit performance indicators at different scales (i.e. network-level, route-level and stop-level) . The proposed platform provides an online prototype for e-science applications in transportation. It allows transit data to be easily accessed, broadly visualized and evaluated. TransitNet offers not only a web-based advanced traveler information system with archived data user service support for data sharing and visualization, but also an interoperable data-rich, regional map–based platform for transportation decision makers and researchers to validate models and existing theories.

7.2 Recommendations for Future Work

Although this study sheds lights on the development of transit performance measures using massive smart card data, more endeavors should be made to enhance both the depth and width of the proposed work. The potential improvements are listed below:

- Bus Dwell Time Analysis

In this study, a fixed time threshold (one hour) is assumed to link different transit trips due to data limitation. Randomness should be taken into account to estimate the bus dwell time (Dueker et al., 2004; Tirachini, 2013), and this will benefit for improving the transit OD estimation accuracy.

- Cloud Computing Technology Integration

Although the transit OD estimation algorithm has been optimized to reduce the computational complexity, its performance is still not satisfying for multiday smart card data. The recently emerging concept and practice in “cloud computing” brings several advantages in terms of data sharing and data processing: scalability, instant cloud hosting, and money saving. Together, with the increasing computing power of IT technology, cloud computing presents a promising solution to large-scale smart card data processing.

- Dynamic Transit Routing

Most transit trip planning systems are either based on the shortest distance or the least number of transfers (Peng and Kim, 2008; Sun et al., 2011). However, an optimal transit itinerary in terms of the distance or number of transfers may not guarantee the least travel time. By integrating transit OD information with E-science based transportation platform, dynamic transit routing function with the shortest travel time can be provided to transit riders.

- Transferability Study

In order to generate broader impacts on other similar AFC systems in other regions, more studies are required to test the effectiveness of the proposed algorithms. Especially for those AFC systems without distance-based fare buses, the applicability of passenger destination estimation algorithm should be further investigated.

Bibliography

US Energy Information Administration, International Energy Outlook 2007. Accessed online at http://www.eia.gov/forecasts/archive/ieo07/index.html, on Nov. 2, 2010.

Li, B., Markov models for Bayesian analysis about transit route origin-destination matrices. *Transportation Research Part B,* 2009, Vol. 43, No. 3, pp. 301-310.

Reddy, A., Lu, A., Kumar, S., Bashmakov, V., Rudenko, S., "Application of entry-only automated fare collection (AFC) system data to infer ridership, rider destinations, unlinked trips, and passenger miles", Preprint CD-ROM for the 88th Annual Meeting of Transportation Research Board, Washington, D.C. 2009.

Hofmann, M., Wilson, S., White, P., "Automated identification of linked trips at trip level using electronic fare collection data", Preprint CD-ROM for the 88th Annual Meeting of Transportation Research Board, Washington, D.C. 2009.

Barry, J.J., Freimer, R., and Slavin, H. "Use of entry-only automatic fare collection data to estimate linked transit trips in New York city", *Transportation Research Record: Journal of the Transportation Research Board*, No. 2112, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 53-61.

Zhao, J., Rahbee, A. and Wilson, N.H.M. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil And Infrastructure Engineering*, 2007, 22, 5,376-387.

Zhang Yu-Fang, Programming on Origin-Destination (OD) matrix estimation-application in New York city mass transit system, Proceedings of the Third International Conference on Traffic and Transportation Studies, 2002, pp. 786-792.

Chen, J., 2009. Research on travel demand analysis of urban public transportation based on smart card data information, Ph.D. dissertation, Tongji University.

Farzin, J. M., "Constructing an automated bus Origin-Destination matrix using farecard and Global Positioning System data in Sao Paulo, Brazil", *Transportation Research Record: Journal of the Transportation Research Board*, No. 2072, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 30-37.

Barry, J.J., Newhouser, R., Rahbee, A., and Sayeda, S. "Origin and destination estimation in New York city with automated fare system data", *Transportation Research Record: Journal of the Transportation Research Board*, No. 1817, Transportation Research Board of the National Academies, Washington, D.C., 2002, pp. 183-187.

Rahbee, A.B. "Farecard passenger flow model at Chicago Transit Authority, Illinois", *Transportation Research Record: Journal of the Transportation Research Board,* No. 2072, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 3-9.

Trépanier, M., Tranchant, N., Chapleau, R., “Individual trip destination estimation in a transit smart card automated fare collection system”, *Journal of Intelligent Transportation Systems*, 11(1):1–14, 2007.

Trépanier, M., Morency, C., Agard, B., “Calculation of transit performance measures using smartcard data”, *Journal of Public Transportation*, 2009, Vol. 12, No. 1.

Nassir, N., Khani A., Lee, S. G., Noh, H., and Hickman, M., Transit stop-level O-D estimation using transit schedule and automated data collection system, Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Zhang, L., Zhao, S., Zhu, Y., and Zhu, Z., Study on the method of constructing bus stops OD matrix based on IC card data. Wireless Communications, Networking and Mobile Computing WiCom 2007, pp. 3147-3150

Seaborn, C., Wilson, N. H. M., Attanucci, J., Using smart card fare payment data to analyze multi-modal public transport journeys (London, UK), Preprint CD-ROM for the 88th Annual Meeting of Transportation Research Board, Washington, D.C. 2009.

Chu, K. K. A. and Chapleau, R., Enriching archived smart card transaction data for transit demand modeling, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2063, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 63-72.

Janssens, D., Wets, W., Brijs, T., Vanhoof, K., Arentze, T., Timmermans, H., “Integrating Bayesian networks and decision trees in a sequential rule-based transportation model”, *European Journal of Operational Research*, 2006, 175, pp. 16-34.

Bayes, Thomas; Price, Mr. "An essay towards solving a problem in the coctrine of chances", *Philosophical Transactions of the Royal Society of London* 53 (0): 370–418, 1763.

Cooper, G. F., “The computational complexity of probabilistic inference using Bayesian belief networks”, *Artificial Intelligence*, Vol. 42, pp. 393-405, 1990.

Arthur, D. and Vassilvitskii, S. k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms. 2007, pp. 1027-1035.

American Public Transportation Association, 2011 Public Transportation Fact Book, 62nd Edition, 2011.

Bagchi, M., and White, P.R. What role for smart-card data from bus systems? Proceedings of the Institution of Civil Engineers: Municipal Engineer, 2004, 157(1), pp.39-46.

Bagchi, M., and White, P.R., The potential of public transport smart card data. *Transport Policy*, 2005, vol. 12, pp.464-474.

Beijing Transportation Research Center, Beijing transportation smart card usage survey, Research Report. 2010.

Beijing Transportation Research Center, Beijing transportation development annual report, Aug. 2011.

Boyle, D. K., Foote, P. J., and Karash, K. H. Public transportation marketing and fare policy. Transportation in the New Millennium, 2000, Accessed on line at: http://onlinepubs.trb.org/onlinepubs/millennium/00093.pdf, on Oct. 7th, 2012.

Cestnik, B. Estimating probabilities: A crucial task in machine learning. In Proceedings of the 9th European conference on artificial intelligence. 1990, pp.147–149. Stockholm.

Chen, J., Research on travel demand analysis of urban public transportation based on smart card data information. Ph.D. dissertation, Tongji University, 2009.

Cheng, C., and Chen, Y., Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*. 2009, vol. 36, pp. 4176-4184.

Chu, K.K. and Chapleau, R. Augmenting transit trip characterization and travel behavior comprehension: multiday location-stamped smart card transactions. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2183, Transportation Research Board of the National Academies, Washington, DC, 2010, pp.29–40.

Chu, K. K., Leveraging data from a smart card Automatic Fare Collection system for public transit planning, Ph.D. dissertation, École Polytechnique De Montréal, 2010.

Cordeiro, R.L.F., Traina, C., Traina, A.J.M., López, J., Kang, U., and Faloutsos, C. Clustering very large multi-dimensional datasets with MapReduce, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. 2011, pp. 690-698.

Cover T. M., Hart, P. E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967, 13 (1): pp.21–27.

Dempsey, P. S. Privacy issues with the use of smart cards. Transit Cooperative Research Program Legal Research Digest 25. 2008. Accessed on line at: http://onlinepubs.trb.org/onlinepubs/tcrp/tcrp_lrd_25.pdf, on Oct. 7th, 2012.

Dinant, J. M., Keuleers, E. Multi-application smart card schemes. Computer Law & Security Report, Vol. 20, no. 1, pp. 22-28. 2004.

Dill, J. Transit use at Transit-oriented developments in Portland, Oregon, area. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2063, Transportation Research Board of the National Academies, Washington, DC, 2008, pp.159–167.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X., A Density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press. 1996, pp. 226-231.

Federal Highway Administration, 2002 status of the nation's highways, bridges, and transit: conditions & performance. Accessed on line at: http://www.fhwa.dot.gov/policy/2002cpr/pdf/execsummary_book.pdf, on Jul. 28th, 2012.

Foote, P. J., Stuart, D. G. and Elmore-Yalch, R., Exploring customer loyalty as a transit performance measure. *Transportation Research Record: Journal of the*

Transportation Research Board, No. 1753, Transportation Research Board of the National Academies, Washington, D.C. 2001, pp. 93-101.

Forgy, E. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biomertrics*, 1965, vol. 21, pp. 768.

Gao, L.X. and Wu, J. P., An algorithm for mining passenger flow information from smart card data, *Journal of Beijing University of Posts and Telecommunications*, Jun. 2011, vol. 34, No.3, pp. 94-97, 2011.

ICF consulting, Center for urban transportation research, Nelson/Nygaard, ESTC. Strategies for Increasing the Effectiveness of Commuter Benefits Programs. TCRP report 87, Transportation Research Board, 2003.

Jang, W, Travel time and transfer analysis using transit smart card data, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2144, Transportation Research Board of the National Academies, Washington, DC, 2010, pp.142–149.

Kitamura, R., T. Yamboardamoto, Y.O. Susilo, and K.W. Axhausen. How routine is a routine? an analysis of day-to-day variability in prism vertex location. *Transportation Research Part A*, 40(3), 2006, pp.259 – 279.

Komorowski, J., Pawlak, Z., Polkowski. L., Skowron, A. Rough sets: A tutorial, in: Pal, S.K., Skowron, A. (Eds.), Rough Fuzzy Hybridization: A New Trend in Decision Making, Springer, Singapore, 1999, pp. 1-98.

Lee, S. G. and Hickman, M., Travel pattern analysis using smart card data of regular users, Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Ma, X., Wang, Y., Feng, C., and Liu, J. Transit smart card data mining for passenger origin information extraction. *Journal of Zhejiang University Science C*, 2012, Vol. 13, No. 10, pp. 750-760.

Mauri, C. Card loyalty. A new emerging issue in grocery retailing. *Journal of Retailing and Consumer Services*. 2003, vol. 10, pp. 13-25.

McKenzie, B. and Rapino, M. Commuting in United States: 2009, American Community Survey Reports. Accessed on line at: http://www.census.gov/prod/2011pubs/acs-15.pdf, on Oct. 7th, 2012.

McGuckin, N. and Nakamoto, Y., Trips, chains, and tours-using an operational definition, Data for Understanding Our Nation's Travel: National Household Travel Survey Conference Transportation, 2004.

Morency, C., M. Trépanier, and B. Agard. Analysing the variability of transit users behaviour with smart card data. The 9th International IEEE Conference on Intelligent Transportation Systems – ITSC 2006, Toronto, Canada, September 17-20, 2006.

Morency, C., M. Trépanier, and B. Agard. Measuring transit use variability with smart-card data, *Transport Policy*, 2007, Volume 14, Issue 3, pp.193-203.

Munizaga, M. A., and Palma, C. Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C*, 2012, Vol. 24, pp. 9-18.

Pawlak, Z., Rough sets. *Informational Journal of Computer and Information Sciences*, 1982, 11(5), pp. 341-356.

Pelletier, M-P., Trépanier, M., Morency, C., Smart card data use in public transit: A literature review, *Transportation Research Part C*, 2011, Vol. 19, Issue 4, pp. 557-568.

Quinlan, J. R. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, 1993.

Rumelhart, D. E., and McClelland, J. L. Parallel distributed processing: explorations in the microstructure of cognition (Vol. 1). 1986, Cambridge, MA: MIT Press.

Skowron, A. and Rauszer, C. The discernibility matrices and functions in information systems, In Intelligent Decision Support: Handbook of Application and Advances of the Rough Sets Theory. Norwell, MA: Kluwer, 1992, pp. 331–362.

Taylor, K. C. and Jones, E. C. Fair fare policies: pricing policies that benefit transit-dependent riders. International Series in Operations Research & Management Science. Vol. 167, Part 3, pp. 251-272, 2012.

Tirachini, A. Estimation of travel time and the benefits of upgrading the fare payment technology in urban transit service. *Transportation Research Part C*, 2012. 10.1016/j.trc.2011.11.007.

Trépanier M., Habib, K. M. N., Morency, C. Are transit users loyal? Revelations from a hazard model based on smart card data, *Canadian Journal of Civil Engineering*. Vol. 39, No. 6, pp. 610-618, 2012.

Utsunomiya, M., Attanucci, J., Wilson, N., Potential uses of transit smart card registration and transaction data to improve transit planning, *Transportation Research Record: Journal of the Transportation Research Board*, No. 1971, Transportation Research Board of the National Academies, Washington, DC, 2006, pp.119-126.

Webb, V., Master Thesis, Customer loyalty in the public transit context, Massachusetts Institute of Technology, 2010.

Wróblewski J. Covering with reducts - a fast algorithm for rule generation. Proc. of RSCTC'98, Warsaw, Poland. Springer-Verlag, Berlin Heidelberg 1998, pp. 402 – 407.

Zhou, T., Zhai C., and Gao Z., Approaching bus OD matrices based on data reduced from bus IC cards. Urban Transport of China, May 2007, vol. 5, no.3, pp. 48-52.

Texas Transportation Institute, 2005 urban mobility report, Texas A&M University, 2005.

Zhang Q., Han, B., and Li, D. Modeling and simulation of passenger alighting anad boarding movement in Beijing metro stations, *Transportation Research Part C: Emerging Technologies,* Volume 16, Issue 5, pp. 635-649, 2008.

Ma, X., Wu, Y., and Wang, Y., DRIVE Net: E-Science transportation platform for data sharing, visualization, modeling, and analysis, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2215, Transportation Research Board of the National Academies, Washington, DC, pp.37-49, 2011.

May, A. D., Traffic Flow Fundamentals, Prentice Hall, Englewood Cliffs, N.J., 1990.

Pipes, L. A., Car-following models and fundamental diagram of road traffic, Transportation Research, Vol. 1, No. 1, 1967, pp. 21–29.

Hey, T., and Trefethen, A. E., The UK e-Science core programme and the grid. Future Generation Computer Systems, Vol. 18, No. 8, 2002, pp. 1017–1031.

Bertini, R. L. and El-Geneidy, A., Generating transit performance measures with archived Data, *Transportation Research Record: Journal of the Transportation Research Board*, No. 1841, Transportation Research Board of the National Academies, Washington, DC, pp.109-119, 2003.

Hamdoucha, Y., Hob, H.W., Sumaleeb, A., and Wangb, G., Schedule-based transit assignment model with vehicle capacity and seat availability, *Transportation Research Part B: Methodological*, Vol. 45, Issue 10, 2011, pp. 1805-1830.

Sun, J., Peng, Z. R., Shan, X., Chen, W., and Zeng, X., Development of web-based transit trip-planning system based on service-oriented architecture, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2217, Transportation Research Board of the National Academies, Washington, DC, pp.87-94, 2011.

Huang, R., and Peng, Z. R., A spatiotemporal data model for dynamic transit networks, *International Journal of Geographical Information Science*, Vol. 22, No.5, pp. 527-545, 2008.

Peng, Z. R. and Huang R., Design and development of interactive trip planning for web-based transit information systems, *Transportation Research Part C*, 2000, Vol. 8, pp.

409-425.

Peng, Z. R., and Kim, E., A Standard-Based Integration Framework for Distributed Transit Trip Planning Systems, *Journal of Intelligent Transportation Systems*, 12(1):13–28, 2008.

Feng, W., Figliozzi, M., Price, S., Feng, W., and Hostetler, K., Techniques to visualize and monitor transit fleet operations performance in urban areas, Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Brian Ferris, OneBusAway: improve the usability of public transit, Doctoral dissertation, University of Washington, 2011.

Kittelson & Associates, Inc., Transit capacity and quality of service manual (1st ed.), TCRP project A-15, Transportation Research Board, National Research Council, Washington, D.C., 1999.

Kittelson & Associates, Inc., Urbitran, Inc. LKC Consulting Services, Inc., Morpace International, Inc., Queensland University of Technology, and Nakanishi, Y., TCRP Report 88, A guidebook for developing a transit performacne-measurement system, Transportation Research Board, National Research Council, Washington, D.C., 2003.

Lem, L. L, Li, J., and Wachs, M., Comprehensive Transit Performance Indicators, 1995 Annual Meeting of the Transportation Research Board, 1995.

Trépanier, M., Morency, C., and Agard, B., Calculation of Transit Performance Measures Using Smartcard Data, *Journal of Pubilc Transportation*, Vol. 12, No. 1, 2009.

Liao, C. and Liu H. X., Development of data-processing framework for transit performance analysis, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2143, Transportation Research Board of the National Academies, Washington, DC, pp.34-43, 2010.

Gallucci, G. and Allen, J. G., regional transit performance measures at Chicago's regional transportation authority, Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Chapleau, R., Chu, K. K., and Allard, B, Synthesizing AFC, APC, GPS and GIS data to generate performance and travel demand indicators for public transit, Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Curries, G., Mesbah, M, Exploring transit operations performance at a network level using AVL and new GIS visualization methods, Preprint CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Lou, Y., Zhang, C., Zheng, Y., Xie Xing, Wang, W., and Huang, Y., Map-matching for low-sampling-rate GPS trajectories, Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 352-361, 2009.

Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxburgh, C., and Byers, A., H., Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011.

Zikopoulos, P., Eaton, C., Deroos, D., Deutsch, T., and Lapis, G., Understanding big data: analytics for enterprise class hadoop and streaming data, Mcgraw Hill, 2011.

Furth, P. G., Hemily, B., Muller, T. H. J., and Strathman, J. G., TCRP report 113: Using archived AVL-APC data to improve transit performance and management, Transportation Research Board, 2006.

Shaw, S. and Rodrigue, J., Geographic Information Systems for Transportation (GIS-T), Accessed online at http://people.hofstra.edu/geotrans/eng/methods/ch1m4en.html, on Nov. 2, 2010.

International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), ISO/IEC 13249-3: Information technology-database languages-SQL multimedia and application packages-part 3: spatial, 2011.

Devillaine, F., Munizaga, M., and Trépanier, M. Detection of activity of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2276, Transportation Research Board of the National Academies, Washington, DC, pp.48-55, 2012.

Lu, A. And Reddy, A. Strategic look at Friday exceptions in weekday schedules for urban transit: Improving service, capturing leisure markets, and achieving cost savings by mining data on automated fare collection ridership. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2274, Transportation Research Board of the National Academies, Washington, DC, pp.30-51, 2012.

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., and Theodoridis, Y. State-of-the-art in Privacy Preserving Data Mining. *ACM SIGMOD Record*, Vol. 33, No. 1, 2004.

Beijing Transportation Research Center, Beijing the 4th Comprehensive Transport Survey Summary Report, Jan. 2012.

Grzymala-Busse, J. and Grzymala-Busse W. An experimental comparison of three rough set approaches to missing attribute values. *Transaction on Rough Set*. Vol. 6, pp.31-50, 2007.

Milligan, G. W. and Cooper, M. C. A study of standardization of variables in cluster analysis. *Journal of Classification* 5: 181-204, 1988.

Vuchic, V. R. *Urban Transit: Operations, Planning and Economics,* John Wiley & Sons, Inc., New York, 2005.

Lyman, K., and Bertini, R. L. Using travel time reliability measures to improve regional transportation planning and operations, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2048, Transportation Research Board of the National Academies, Washington, DC, pp.1-10, 2008.

Federal Highway Administration, Travel Time Reliabiliy: Making it there on time, all the itme, 2006. Accessed on line at: http://ops.fhwa.dot.gov/publications/tt_reliability/, on Apr. 18th, 2013.

Chiu, Y. C., and Mirchandani, P. B., Online Behavior-Robust Feedback Information Routing Strategy for Mass Evacuation. IEEE Transactions on Intelligent Transportation Systems. Vol. 9, No. 2, 264-274. 2008.

Murray-Tuite, P and Mahmassani, H. (2005). Identification of Vulnerable Transportation Infrastructure and Household Decision Making Under Emergency Evacuation Conditions. Research Report for Southwest Region University Transportation Center, SWUTC/05/167528-1.

May, A., Traffic Flow Fundamentals, Prentice Hall, Englewood Cliffs, NJ, 1990.

Pipes, L., A. Car Following Models and Fundamental Diagram of Road Traffic, *Transportation Research*, vol. 1, no. 1, 1967.

Smith, MC, Sadek, AW, and Huang, S. Large-Scale Microscopic Simulation: Toward an Increased Resolution of Transportation Models. *Journal of Transportation Engineering*, Vo. 134, No. 7, 273-281. 2008.

Haghani, A., Tian, Q, Hu, H. Simulation Model for Real-Time Emergency Vehicle Dispatching and Routing. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1882, Transportation Research Board of the National Academies, Washington, DC, pp.176-183, 2008.

Hey, T. and Trefethen, A.E., The UK e-Science Core Programme and the Grid. *Future Generation Computer Systems*, Vol. 18, Issue 8, 1017 – 1031, 2002.

ESRI, ESRI Shapefile Technical Description: An ESRI While Paper, 1998, Accessed on line at: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf, on Apr. 18th, 2013.

Wu, Y.-J., and Wang, Y., An Interactive Web-based System for Urban Traffic Data Analysis, *International Journal of Web Applications*, Vol. 1, No 4, 2009, pp. 241-252.

Wu, Y.-J., Wang, Y., and Qian, D., A Google-Map-Based Arterial Traffic Information System, IEEE International Conference on 10th Intelligent Transportation Systems, Seattle, Washington, USA, 2007.

National Transportation Communications for ITS Protocol, Accessed on line at: http://www.ntcip.org/library/documents/pdf/ap-datex_980102_w2.pdf on Jul. 20th, 2010.

Council, F. M. and Mohamedshah, Y. M., Highway Safety Information System Guidebook for the Washington State Data Files, 2009, http://www.hsisinfo.org/guidebooks/WA_Guidebook_2009.htm, Accessed July 02, 2010.

Wang, Y., Corey, J., Lao Y., and Wu, Y.-J., Development of a Statewide Online System for Traffic Data Quality Control and Sharing, Transportation Northwest (TransNow), Project 61-6022, 2009.

Ma, X., McCormack, E., Wang, Y., Process commercial GPS data to develop a web-based truck performance measures program, Paper 11-1932, Proceedings CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C. 2011.

Malinovskiy, Y., Wu, Y.-J., Wang, Y., Lee, U., Field Experiments on Bluetooth-based Travel Time Data Collection, Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C. Paper Number: 11-3056, 2011.

Codd, E.F.. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM,* 13 (6): 377–387. 1970.

Ran, B., Chang, B. P. and Chen, J., Architecture Development for Web-based Geographic Information System Applications in Transportation, *Transportation Research Record: Journal of the Transportation Research Board*, No. 1660, Transportation Research Board of the National Academies, Washington D.C. 1999, pp. 114-121.

Eckerson, W. W., Three Tier Client/Server Architecture: Achieving Scalability, Performance, and Efficiency in Client Server Applications. Open Information Systems 10, 1995.

Rahm, E. and Do, H. H., Data Cleaning: Problems and Current Approaches. *IEEE Techn. Bulletin on Data Engineering*, Dec. 2000

Wu, Y.-J., G. Zhang, and Y. Wang. Volume Data Correction for Single-Channel Advance Loop Detectors at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2160, Transportation Research Board of the National Academies, Washington D.C. 2010, pp. 128-139.

Cheevarunothai, P., Y. Wang, and N. L. Nihan. Identification and Correction of Dual-Loop Sensitivity Problems. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1945, Transportation Research Board of the National

Academies, Washington D.C. 2006, pp. 73-81.

Garrett, J.J., Ajax: A New Approach to Web Applications, http://www.adaptivepath.com/publications/essays/archives/000385.php, Accessed on July 02, 2010.

Wu, Y.-J., An, S., Ma, X. and Wang, Y., Development of a Web-based Arterial Network Analysis System for Real-time Decision Making, *Transportation Research Record: Journal of the Transportation Research Board*. No. 2215, 2011, pp. 24-36.

Yu, R., Lao, Y., Ma, X. and Wang, Y., Short-Term Traffic Flow Forecasting for Improved Estimates of Freeway Incident Induced Delays, Paper 11-3593, Proceedings CD-ROM for the 90th Annual Meeting of Transportation Research Board, Washington, D.C., 2011.

Google, Inc. The Google Map API. http://code.google.com/apis/maps/index.html, Accessed on July 02, 2010.

Google, Inc., Google Chart Tools, http://code.google.com/p/gwt-google-apis/wiki/VisualizationGettingStarted, accessed on July 02, 2010.

Google, Inc., Google Web Toolkit", http://code.google.com/webtoolkit/, Accessed on July 02, 2010

Eclipse Foundation, http://www.eclipse.org/, Accessed on July 02. 2010.

The PostgreSQL Global Development Group, PostgreSQL 9.2.4 Documentation, http://www.postgresql.org/files/documentation/pdf/9.2/postgresql-9.2-A4.pdf, accessed on April 02, 2013.

Refractions Research, PostGIS 2.0 Manual, http://postgis.net/stuff/postgis-2.0.pdf, accessed on April 02, 2013.

Vaadin Ltd., Book of Vaadin: 4th Edition, 2012, https://vaadin.com/download/book-of-vaadin/vaadin-6/pdf/book-of-vaadin.pdf, accessed on April 02, 2013.

Dueker, K.J., Kimpel, T.J., Strathman, J. G., and Callas, S., Determinants of Bus Dwell Time, *Journal of Public Transportation,* Vol. 7, Issue 1, 2004.

Tirachini, A., Bus dwell time: the effect of different fare collection systems, bus floor level and age of passengers, *Transportmetrica A: Transport Science*, Vol. 9, No. 1, pp. 28–49, 2013.

Curriculum Vitae

EDUCATION

Doctor of Philosophy in Transportation Engineering……………………Jun 2013(Expected)
University Of Washington, Seattle, WA,
Department of Civil and Environmental Engineering

Certificate in Global Trade, Transportation, and Logistics Studies.......................June 2010
University of Washington, Seattle, WA,
The Global Trade, Transportation, and Logistics (GTTL) Studies Program

Master of Science in Transportation Engineering………………………………Mar 2010
University Of Washington, Seattle, WA,
Department of Civil and Environmental Engineering

Bachelor of Engineering in Electronic Engineering………………………………July 2007
Beijing Institute of Technology, Beijing, China,
Department of Electronic Engineering

RESEARCH INTERESTS

- Big data in transportation
- Large-scale transportation data mining and management
- Public transit
- Intelligent Transportation Systems (ITS)
- Traffic operations and simulation
- Freight and logistics

HONORS

PacTrans Travel Awards	2012
Boeing/GTTL Academic Achievement Awards	2010
NACOTA Publication Fellowship	2010
NACOTA Logistic Fellowship	2010
Annual Excellent Student Award, Beijing Institute of Technology	2003~2007
Outstanding Graduate Award, Beijing Institute of Technology	2007

JOURNAL PUBLICATIONS

Accepted:

1. **Xiaolei Ma**, Yao-Jan Wu, and Yinhai Wang. "DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis", *Transportation Research Record: Journal of the Transportation Research Board.* Vol. 2215, pp.37-49, 2011.

2. **Xiaolei Ma,** Edward McCormack and Yinhai Wang. “Processing Commercial GPS Data to Develop a Web-Based Truck Performance Measures Program”, Accepted for *Transportation Research Record: Journal of the Transportation Research Board.* Vol. 2246, pp.24-36, 2011.

3. Yao-Jan Wu, An Shi, **Xiaolei Ma**, and Yinhai Wang. “Development of a Web-based Arterial Network Analysis System for Real-time Decision Making”, *Transportation Research Record: Journal of the Transportation Research Board*. Vol. 2215, pp.92-100, 2011.

4. **Xiaolei Ma**, Yinhai Wang, Feng Chen and Jianfeng Liu. "Transit smart card data mining for passenger origin information extraction", *Journal of Zhejiang University Science C (SCI)*, Vol. 13, No. 10, pp. 750-760, 2012.

5. Yong Wang, **Xiaolei Ma**, Yunteng Lao, Yinhai Wang, and Haijun Mao, “Location Optimization of Multiple Distribution Centers Under Fuzzy Environment”, *Journal of Zhejiang University Science A*, Vol.13, No. 10, pp. 782-798, 2012.

6. Jianyang Zheng, **Xiaolei Ma**, Yao-Jan Wu, Yinhai Wang, "Measuring Signalized Intersection Performances in Real Time with Traffic Sensors", *Journal of Intelligent Transportation Systems*, DOI:10.1080/15472450.2013.771105, 2013.

7. Runze Yu, Yunteng Lao, **Xiaolei Ma**, Yinhai Wang, “Short-Term Traffic Flow Forecasting for Freeway Incident Induced Delays”, *Journal of Intelligent Transportation Systems*, In Press, 2013.

8. Yong Wang, **Xiaolei Ma**, Yunteng Lao, Yinhai Wang, and Haijun Mao, “Vehicle Routing Problem: Simultaneous Deliveries and Pickups with Split Loads and Time Windows”, *Transportation Research Record: Journal of the Transportation Research Board*, In Press, 2013.

Under Review:

9. **Xiaolei Ma**, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu, “Mining Smart Card Data for Transit Riders’ Travel Patterns”, *Transportation Research Part C: Emerging Technologies*, Under review.

10. Yong Wang, **Xiaolei Ma**, Yunteng Lao, Yinhai Wang and Haijun Mao, "A Fuzzy-based Customer Clustering Approach with Hierarchical Structure for Logistics Network Optimization", *Expert Systems with Applications*, Under review.

11. Yong Wang, **Xiaolei Ma**, Yunteng Lao, Yinhai Wang and Haijun Mao, "A Two-stage Heuristic Method for Vehicle Routing Problem with Split Deliveries and Pickups", *Annals of Operations Research*, Under review.

12. Yong Wang, **Xiaolei Ma**, Yunteng Lao, Yinhai Wang and Haijun Mao, "A Hybrid Algorithm for Two-echelon Logistics Distribution Region Partitioning Problem", *Annals of Operations Research*, Under review.

CONFERENCE PROCEEDINGS

1. **Xiaolei Ma**, Yao-Jan Wu, Yinhai Wang, Feng Chen, and Jianfeng Liu, "Mining Smart Card Data for Transit Riders' Travel Patterns", *Preprint CD-ROM, the 92nd Annual Meeting of the Transportation Research Board,* Washington, D.C., Jan. 2013.

2. Yong Wang, **Xiaolei Ma**, Yunteng Lao, Yinhai Wang, and Haijun Mao, "Vehicle Routing Problem: Simultaneous Deliveries and Pickups with Split Loads and Time Windows", *Preprint CD-ROM, the 92nd Annual Meeting of the Transportation Research Board,* Washington, D.C., Jan. 2013.

3. **Xiaolei Ma**, Runze Yu, Yinhai Wang, "Developing a Regional Map-Based Platform for Spatial and Temporal Assessment of Traffic Emission Inventory", *Proceedings of COTA International Conference of Transportation Professionals (CICTP)*. Aug. 2012.

4. **Xiaolei Ma**, Yao-Jan Wu, and Yinhai Wang. "DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis", *Preprint CD-ROM, the 90th Annual Meeting of the Transportation Research Board*, Washington, D.C., Jan. 2011.

5. **Xiaolei Ma**, Edward McCormack and Yinhai Wang. "Processing Commercial GPS Data to Develop a Web-Based Truck Performance Measures Program", *Preprint CD-ROM, the 90th Annual Meeting of the Transportation Research Board*, Washington, D.C., Jan. 2011.

6. Yao-Jan Wu, An Shi, **Xiaolei Ma**, and Yinhai Wang. "Development of a Web-based Arterial Network Analysis System for Real-time Decision Making", *Preprint CD-ROM, the 90th Annual Meeting of the Transportation Research Board*, D.C., Jan. 2011.

7. Runze Yu, Yunteng Lao, **Xiaolei Ma**, Yinhai Wang, "Short-Term Traffic Flow Forecasting for Improved Estimates of Freeway Incident Induced Delays", *Preprint CD-ROM, the 90th Annual Meeting of the Transportation Research Board*, D.C., Jan. 2011.

8. **Xiaolei Ma**, Guohui Zhang, Jonathan Corey, Yinhai Wang, "Simulation-based Investigations on Transit Signal Priority System Operations under Coordinated Control Strategy", *Proceedings of the 10th International Conference of Chinese Transportation Professionals (the 10th ICCTP).* Aug. 2010. (**NACOTA Publication and Logistic Fellowship**)

9. **Xiaolei Ma**, Edward McCormack, "Using Truck Fleet Management GPS Data to Develop the Foundation for a Performance Measures Program", *Preprint CD-ROM, the 89th Annual Meeting of the Transportation Research Board*, Washington, D.C., Jan. 2010.

10. Jianyang Zheng, **Xiaolei Ma**, Yinhai Wang, Ping Yi, "Measuring Signalized Intersection Performances in Real-Time with Traffic Sensors", *Preprint CD-ROM, the 88th Annual Meeting of the Transportation Research Board,* Washington, D.C., Jan. 2009.

PRESENTATIONS

1. **Xiaolei Ma**, Yinhai Wang, Feng Chen and Jianfeng Liu. "Transit smart card data mining for passenger origin information extraction", Presented at 92nd Annual Meeting of the

Transportation Research Board, Washington, D.C., Jan. 2013.

2. **Xiaolei Ma**, "*DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis*", Presented at CEE 500 graduate seminar in University of Washington, Apr. 2012.

3. **Xiaolei Ma**, Yao-Jan Wu, and Yinhai Wang. "*DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis*", Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 2011.

4. **Xiaolei Ma**, Guohui Zhang, Jonathan Corey, Yinhai Wang, "*Simulation-based Investigations on Transit Signal Priority System Operations under Coordinated Control Strategy*", Presented at the 10th International Conference of Chinese Transportation Professionals (the 10th ICCTP). Aug. 2010.

5. Wang, Yinhai, Yao-Jan Wu, Jonathan Corey, and **Xiaolei Ma**. "*Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net).*" Presented at North American Travel Monitoring Exposition and Conference (NATMEC). Seattle, WA. Jul. 2010.

6. Junfeng Jiao, **Xiaolei Ma**, Sunny Rose, Erica Wygonik, *E-Grocery: Transportation Sustainability and Market Potential*, Presented at Global Trade, Transportation, and Logistics Studies, 16th Annual Conference, University of Washington, Jun. 8th, 2010

7. Yao-Jan Wu, **Xiaolei Ma** and Yinhai Wang, *Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net): Prototype Development and Implementation.* Presented at Puget Sound Regional Council (PSRC) Regional Traffic Operators Committee (RTOC) Meeting, Apr. 1st, 2010

PROJECT REPORTS

1. Mark Hallenbeck, Yinhai Wang, John Ishimaru, **Xiaolei Ma**, Mike Richards, Jonathan Corey. "*Clean-Up of Existing Data Set to Support Dynamic Mobility Applications Development*". Project Report for U.S. Department of Transportation and Federal Highway Administration. DTFH61-11-H-00026, Dec. 2011

2. Yinhai Wang, Yao-Jan Wu, **Xiaolei Ma**, Jonathan Corey. "*Real-time Travel Time Prediction on Urban Traffic Network*", Project Report for Transportation Northwest (TransNow). No. 61-7212, Oct. 2010.

3. Edward McCormack, **Xiaolei Ma**, Charles Klocow, Anthony Currarei, Duane Wright. "*Develop a GPS-based Truck Performance Measure Platform*", Project Report for Washington State Department of Transportation and Transportation Northwest (TransNow), WA-RD 748.1, Apr. 2010.

4. Yinhai Wang, Mark Hallenbeck, Jianyang Zheng, Guohui Zhang, **Xiaolei Ma**, and Jonathon Corey. "*Comprehensive Evaluation on Transit Signal Priority System Impacts Using Field Observed Traffic Data*". Project Report for Washington State Department of Transportation and Transportation Northwest (TransNow), WA-RD 699.1/TNW2008-06, Jun. 2008.

RESEARCH PROJECTS

Migration of Operations Datamart to DRIVE Net

Washington State Department of Transportation……………………......Jul 2011~Jun 2013

- Developed a geospatial database to store the roadway network, and link each roadway segment with transportation data
- Developed an E-Science of transportation platform for data sharing, visualization, modeling, and analysis

Beijing Transit Data Mining for Passenger Origin-Destination Information Extraction

Beijing Transportation Research Center……… ……………………….Sept 2010~Jul 2012

- Proposed a GPS-based boarding information extraction algorithm
- Developed a Markov chain based Bayesian decision tree for passenger origin inference
- Developed a DBSCAN clustering algorithm for travel pattern discovery
- Developed a K-Means++ and Rough Set theory procedure for regularity extraction
- Developed a transfer-based analysis and historical pattern recognition procedure for passenger destination inference
- Developed a Visual C# based Beijing Transit Data Mining Tool software

Clean-up of Existing Data Sets to Support Dynamic Mobility Applications Development

U.S. Department of Transportation and Federal Highway Administration…Jul 2011~Dec 2011

- Developed a stand-alone Openlayers based map interface for heterogeneous transportation data fusion and visualization

Quantifying Incident-Induced Travel Delays on Freeways Using Traffic Sensor Data: Phase II

Washington State Department of Transportation (WSDOT…………….Mar 2010~Aug 2010

- Implemented a web-based Incident Induced Delay calculation module using time series technique

Real-time Travel Time Prediction on Urban Traffic Network

Washington State Department of Transportation (WSDOT)…….…….Jun 2009~Nov 2010

- Developed a real-time dynamic shortest travel time route identification algorithm
- Developed a Kalman Filter algorithm to predict the arterial travel time

GPS based Truck Freight Performance Measure Platform

Washington State Department of Transportation (WSDOT)…………...Jul 2008~Mar 2010

- Developed a freight database and GPS data collection system
- Developed an automatic trip identification algorithm to detect truck's origin and destination
- Developed an automatic freight performance measures generation algorithm
- Developed a web-based freight performance measure visualization platform

Development of a Statewide Online System for Traffic Data Quality Control and Sharing

Washington State Department of Transportation (WSDOT)……………..Aug 2009~Oct 2009

- Developed a web-based loop data retrieval and analysis module.

Comprehensive Evaluation of Transit Signal Priority System Impacts Using Field Observed Traffic Data

Washington State Department of Transportation (WSDOT)………….Dec 2007~Jun 2008

- Improved the Transit Signal Priority (TSP) algorithm, and implemented the algorithm using Vehicle Actuated Programming (VAP)
- Simulation and test in VISSIM
- Data processing and analysis

TEACHING EXPERIENCE

Teaching Assistant, CEE 412/599, Transportation Data Management……Winter Quarter, 2013
Department of Civil and Environmental Engineering, University of Washington, Seattle

Pre-doctoral Instructor, CEE 590, Transportation Systems Operation Autumn Quarter, 2012
Department of Civil and Environmental Engineering, University of Washington, Seattle

- 19 students are enrolled
- Taught twice per week (1 hour and 20 minutes)
- Designed and conducted 6 lab sessions, 6 assignments, 2 midterms, and 3 projects
- Teaching evaluation score is 4.1/5 (Very Good)

Teaching Assistant, CEE 590, Transportation Systems Operation….Autumn Quarter, 2009
Department of Civil and Environmental Engineering, University of Washington, Seattle

WORK EXPERIENCE

Beijing Transportation Research Center…………………………………Sept 2010~Aug 2012

- Constructing transit origin-destination matrix using smartcard and GPS data in Beijing

China Telecom Group Beijing Corporation………………………………Apr 2007~Jun 2007

- Internship in the Network Department, Information Service Centre, participated in the project of Mega Eyes and Videoconference system

Beijing Chang Cheng Aeronautical M&C Technology Co., Ltd………….Jun 2006~Jul 2006

- Participate in developing an underground staff positioning system project

PROFESSIONAL ACTIVITIES

Paper Reviews

- International Conference of Chinese Transportation Professionals
- Annual Meeting of the Transportation Research Board
- International IEEE Conference on Intelligent Transportation Systems
- Expert Systems with Applications
- Public Transport

Editorial Committee Member:

- The 9th International Conference of Chinese Transportation Professionals

AFFILIATIONS

Institute of Transportation Engineers, Student member……...............................2008~Present
American Society of Civil Engineers (ASCE), Student member………………...2009~Present

North American Chinese Overseas Transportation Association (NACOTA), Student member………………………………………………………………………….2009~Present
IEEE student member……………………………………………………….................2012~Present

COMPUTER SKILLS

- *Programming Language:* Java (proficient in J2EE framework), C#, C/C++, PHP, Python
- *Programming Platform:* Microsoft Visual Studio, Eclipse (proficient with Google Web Toolkits)
- *Operating Systems:* Windows, MacOS, Linux, DOS
- *Database:* Microsoft SQL Server, PostgreSQL, MYSQL, Microsoft Access
- *Data Processing and Analysis:* R, MATLAB, EXCEL, SPSS
- *Transportation Simulation:* VISSIM, Synchro
- *GIS Software:* ArcGIS, PostGIS, Geoserver

Reprint Permissions

Permissions have been granted by the publishers to reuse the contents in the papers listed below in this dissertation.

Xiaolei Ma, Yinhai Wang, Feng Chen and Jianfeng Liu, "Transit Smart Card Data Mining for Passenger Origin Information Extraction", *Journal of Zhejiang University Science C (SCI)*, Vol. 13, No. 10, pp.750-760. 2012

Xiaolei Ma, Yao-Jan Wu, and Yinhai Wang, "DRIVE Net: An E-Science of Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis", *Transportation Research Record: Journal of the Transportation Research Board*, Vol.2215, pp.37-49, 2011.

Xiaolei Ma, Yao-Jan Wu, Feng Chen, Jianfeng Liu and Yinhai Wang, "Mining Smart Card Data for Transit Riders' Travel Patterns", *Preprint CD-ROM, the* 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 2013.