The Planning and Analysis Implications of Automated Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modeling Examples

By

Jinhua Zhao

Bachelor of Engineering in Urban Planning, Tongji University (2001)

Submitted to the Department of Urban Studies and Planning and the Department of Civil and Environmental Engineering in Partial Fulfillment of the Requirements for the degrees of

Master in City Planning

and

Master of Science in Transportation

at the

Massachusetts Institute of Technology

September 2004

Signature of Author .....

Department of Urban Studies and Planning
June 2004

Certified by.....

Nigel H.M. Wilson
Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by.....

Joseph Ferreira, Jr.
Professor of Urban Studies and Operations Research
Thesis Supervisor

Accepted by.....

Dennis M. Frenchman
Professor of the Practice of Urban Planning
Chairman, Master in City Planning Committee

Accepted by.....

Nigel H.M. Wilson
Professor of Civil and Environmental Engineering
Chairman, Master of Science in Transportation

The Planning and Analysis Implications of Automated Data Collection Systems: Rail Transit OD Matrix Inference and Path Choice Modeling Examples

By
Jinhua Zhao

Submitted to the Department of Urban Studies and Planning and the Department of Civil and Environmental Engineering in Partial Fulfillment of the Requirements for the degrees of Master in City Planning and Master of Science in Transportation at the Massachusetts Institute of Technology September 2004

Abstract:

Transit agencies in U.S. are on the brink of a major change in the way they make many critical planning decisions. Until recently transit agencies have lacked the data and the analysis techniques needed to make informed decisions in both long-term planning and day-to-day operations. Now these agencies are entering an era in which a large volume of raw data will be available due to the implementation of ITS technology including Automated Data Collection systems (ADC), such as Automated Fare Collection systems (AFC), Automated Vehicle Location systems (AVL), and Automatic Passenger Counting systems (APC).

Automated Data Collection systems have distinct advantages over the traditional data collection methods: large temporal and spatial coverage, continuous data flow and currency, low marginal cost, accuracy, automatic collection and central storage, etc. Thanks to these unique features, there exists a great potential for ADC systems to be used to support decision-making in transit agencies. However, effectively utilizing ADC systems data is not straightforward. Several examples are given to illustrate that there is a critical gap between what ADC systems directly offer and what is needed practically in public transit agencies' decision-making practice. Meanwhile, the framework of data processing and analysis is not readily available, and transit agencies generally lack needed qualified staff. As a consequence, these data sources have not yet been effectively utilized in practice.

A strong foundation of ADC data manipulation, analysis methodologies and techniques with support of advanced technologies such DBMS and GIS is required before the full value of the new data source can be exploited. This research is an initial attempt to lay

out such a framework by presenting two case studies both in the context of the Chicago Transit Authority.

One study proposes an enhanced method of inferring the rail trip OD matrix from an origin-only AFC system to replace the routine passenger survey. The proposed algorithm takes advantage of the pattern of a person's consecutive transit trip segments. In particular the study examines the rail-to-bus case (which is ignored by prior studies) by integrating AFC and AVL data and utilizing GIS and DBMS technologies. A software tool is developed to facilitate the implementation of the algorithm.

The other study is of rail path choice, which employs the Logit and Mixed Logit models to examine revealed public transit riders' travel behavior based on the inferred OD matrix and the transit network attributes. This study is based on two data sources: the rail trip OD matrix inferred in the first case study and the attributes of alternative paths calculated from a network representation in Trans CAD. This study demonstrates that a rigorous traveler behavior analysis can be performed based on the data source from ADC systems.

Both cases illustrate the potential as well as the difficulty of utilizing these systems and more importantly demonstrate that at relatively low marginal cost, ADC systems can provide transit agencies with a rich information source to support decision making. The impact of a new data collection strategy, and effective data analyses and utilization may affect transit agency decision making practice and push them into another stage wherein decisions are supported by an information-rich system based on more comprehensive data sources and more effective data analysis.

Thesis Supervisor:	Nigel H.M. Wilson
Title:	Professor of Civil and Environmental Engineering
	Department of Civil and Environmental Engineering
Thesis Supervisor:	Joseph Ferreira, Jr.
Title:	Professor of Urban Studies and Operations Research
	Department of Urban Studies and Planning

Acknowledgement

One can work on his thesis for months and years. However, it is some critical moments of joy that make those efforts worthwhile. In the past one year, the moments to me were every Thursday afternoon when I was in Professor Nigel Wilson's office. He made me appreciate the positive and the negative beauties of research, both of which were challenging me the most. I walked out of his office with refilled confidence and passion. Nigel is my Mentor.

Professor Joseph Ferreira has guided me through my life at MIT for three years. It is he who encouraged a newcomer to become a confident MIT member, who enabled me to present myself in front of people, and who invited me, the "homeless" international, to his family Thanksgiving dinner.

I would like to thank Professor Frank Levy for reading my paper and cautioning me about "the potential mismatch between a quite fine-grained model and a data set that is fairly coarse" and "not adding a lot of refinements which extend the model's precision beyond what the data can support". I thank Mikel Murga for his instruction and suggestion on transportation network modeling and his inspiring remark "a bit thinking before the working".

Thanks to Adam Rahbee at CTA who supervised me when I was interning in CTA for initiating the project that later evolved into this thesis and all his great help and advice throughout the research. Thanks to Kevin O'Melly and Mike Heynes who provided me the AFC and AVL data, and to Jeff Sriver for his help on the Circle Line project information.

Thanks to Michel Bierlaire for developing the BIOGEME software for model estimation and answering dozens of my questions. Thanks to Virginia Siggia, Michael Segal and Martha Tai for reading my drafts and correcting tons of grammatical errors.

A special thank you goes to Jiang Zhanbin, Martha Tai, Guo Zhan and Guo Ming, my best friends in the U.S., and Tan Zhengzhen, my beloved girlfriend.

The heartfelt thanks to my parents Zhao Jingwei and Jiang Huiling, and my sister Zhao Yu for their love. This thesis is dedicated to them.

Contents

List of Figures

List of Tables

Chapter One: Introduction

Transit agencies in U.S. are on the brink of a major change in the way they make many critical planning decisions. Until recently transit agencies have lacked the data and the analysis techniques needed to make informed decisions in both long-term planning and day-to-day operations. Now these agencies are entering an era in which a large volume of raw data will be available due to the implementation of ITS technology including Automated Data Collection systems (ADC), such as the Automated Fare Collection system (AFC), the Automated Vehicle Location system (AVL), and the Automatic Passenger Counting system (APC).

This research examines characteristics of these ADC systems, illustrates the potential as well as the difficulty of utilizing these systems, argues that a framework of ADC data manipulation and analysis methodologies and techniques is required before the full value of ADC systems can be realized, and demonstrates that at relatively low marginal cost, ADC systems can provide transits agencies with very rich information to support their critical decision making processes.

1.1 ADC systems: Potentials and Reality

Automated Data Collection systems have distinct advantages over the traditional data collection methods: large temporal and spatial coverage, continuous data flow and currency, low marginal cost, accuracy, automatic collection and central storage, etc.

Thanks to these unique features, there exists a great potential for ADC systems to be used to support decision-making in transit agencies. However, effectively utilizing ADC systems data is not straightforward. This study offers examples to illustrate that there is a critical gap between what ADC systems directly offer and what is needed practically in public transit agencies' decision-making practice.

Furthermore, inherent disadvantages of the ADC system data make the gap even harder to bridge. Firstly, many of the ADC systems are not primarily designed for data collection; data collection is just an auxiliary function and so some critical information may be lacking. The data collected is often intermittent, fragmented, and not in an easy-to-use format, which makes the raw data unintelligible and meaningless without intensive processing and careful interpretation. Secondly, different systems are often implemented separately by different vendors. The data structures and formats are incompatible; the data are stored and managed in different database management systems; the ADC systems are implemented at different locations and in different time periods; all of this makes the integration across systems very difficult and seriously limits the utilization of the ADC system data.

Meanwhile, a framework for data processing and analysis methodologies and techniques for ADC data is not readily available, and transit agencies generally lack needed qualified staff. As a consequence, these data sources are seldom, if ever, effectively utilized in practice. A huge amount of data is being produced day after day but most of it is not being organized into information of interest to transit agencies.

1.2 Research Objectives

The primary objective of the study is to examine how to effectively utilize ADC systems to support public transit agencies' decision making by setting up a framework including methodologies and techniques for ADC data manipulation and analysis with the support of advanced technologies such as DBMS and GIS.

This primary objective can be divided into three subordinate objectives: the first is to examine the common characteristics of ADC systems in general and two specific systems currently operational in the Chicago Transit Authority (CTA), with the objective of illustrating the potential and the difficulty of using ADC data, and demonstrating the gap between the data available from ADC systems and the information useful to transit agencies.

The second is to develop an improved methodology for rail origin-destination matrix inference by integrating the AFC and AVL systems, develop a software tool to automate this methodology, and practically implement it in the context of the CTA rail system.

The third is to develop a discrete choice model to characterize public transit riders' rail path choice behavior, to estimate the path choice model using data on the inferred CTA rail OD matrix and path attributes calculated from a TransCAD network representation, and to apply the model to assess the impact on the rail path's market share of a proposed network change.

1.3 ADC Data Manipulation and Analysis Framework

The possibility and value of setting up a framework of ADC data manipulation and analysis are illustrated through a series of examples and case studies involving processing and interpretation of the ADC systems data.

Three short examples are presented here to show the techniques to process the raw ADC records to make basic sense of the data: these are, processing AFC data to obtain the bus ID, processing AVL data to get the bus stop ID, and integrating AFC and AVL data to relate passenger trips to vehicle trips in order to obtain the passenger boarding bus stop ID. Although the bus ID and the boarding bus stop ID are essential information needed to interpret the data, neither of them are readily provided by the ADC systems. Difficult and cumbersome data processing is involved to transform "data" into "information". This process requires familiarity with the structure of the ADC systems, knowledge of data manipulation and analysis technologies, as well as an understanding of transportation planning.

Therefore a framework covering methodologies and techniques for the ADC data manipulation and analysis, with the support of advanced technologies such as DBMS and GIS, is indispensable in order to exploit the full potential of the new ADC system data and bridge the gap between what's becoming available from ADC systems and what's needed by transit agencies.

This research is an initial attempt to lay out such a framework by presenting two case studies, both in the context of the Chicago Transit Authority (CTA). One study involves integrating the AFC and AVL data to infer the rail origin-destination (OD) matrix from the origin-only rail trip data. The other study is of rail path choice, which employs discrete choice models to examine revealed public transit riders' travel behavior based on the inferred OD matrix and transit network attributes.

Each case entails its own set of methodologies. In particular, the examination and characterization of passengers' trip segment chains, the integration of the AFC and AVL systems, DBMS, GIS and programming techniques are involved in the rail OD inference application. Discrete choice analysis methods particularly the Logit and Mixed Logit models, network representation and analysis in TransCAD are used in the path choice modeling application.

The three ADC data processing examples and two in-depth case studies involve different scales and complexity, use different specific techniques, and solve different problems. However, they all utilize data from ADC systems, they all aim to support certain aspects of transit agencies' decision making, and they are designed within the framework of data manipulation and analysis methodologies and techniques that facilitates the effective utilization of ADC systems.

1.4 Three Stages of Transit Agency Decision Making Practice in Terms of Data Support

The relationship between transit agency decision-making and data analysis support can be categorized into three stages. Traditionally, transit agencies have been painfully short of data to support their decision-making processes. Traditional data collection strategies are expensive and as a result are carried out only infrequently, which cannot support rigorous research and analyses that require timely and comprehensive data. This stage of development of transit agency decision-making is termed "Stage I" as in Figure 1-1.

Applying new technology to automated data collection has been an important trend in data collection. ADC systems are changing transit agencies from being data poor to data rich. Although huge amounts of data are now being continuously produced, however, they have not typically been transformed into information useful to transit agencies. This stage of development is referred to as Stage II in Figure 1-1.

Figure 1-1 Three Stages of Transit Agency Decision Making Practice in Terms of Data Support

Stage	Rich Data Sources	Effective Data Utilization to Support Decision Making
I	No	No
II	Yes	No
III	Yes	Yes

Source: the author. All following figures and tables are created by the author unless noticed otherwise.

This research illustrates the potential as well as the difficulty of utilizing these systems, offers an initial attempt to lay out a framework for ADC data manipulation and analysis

and demonstrates that at relatively low marginal cost, ADC systems can provide transit agencies with a rich information source to support decision making. The impact of a new data collection strategy, and effective data analyses and utilization may affect transit agency decision making practice and push them into another stage wherein decisions are supported by an information-rich system based on more comprehensive data sources and more effective data analysis. This is referred to as Stage III in Figure 1-1.

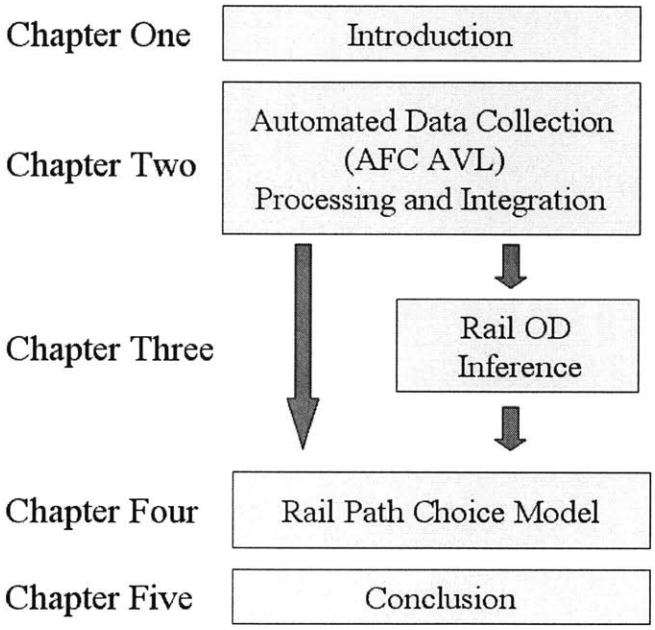
1.5 Thesis Organization

This thesis is organized into five chapters as shown in Figure 1-2. Chapter Two examines the common characteristics of ADC systems and introduces two specific systems currently operational in CTA. Three examples of preliminary ADC data processing are presented to illustrate the potential and obstacles associated with better ADC data utilization.

Chapters Three and Four present two in-depth case studies of effective utilization of ADC data. Chapter Three proposes an improved method of inferring the rail trip OD matrix from an origin-only system by taking advantage of data integration across multiple ADC systems. The proposed method could partially replace the routine passenger survey with respect to estimating rail OD matrices. A software tool is developed to automate the destination inference process so as to make this method practical. The software is then applied to the CTA rail system to generate a one-week OD matrix, and the estimated trip segment chain patterns are analyzed.

Chapter Four examines public transit riders' path choice behavior in the rail system. The path choice decision is modeled using Multinomial Logit (MNL) and Mixed Logit model forms, utilizing data from the OD matrix estimated in chapter three and the path attributes determined from the CTA rail network represented in TransCAD. The specification and interpretation of a series of models are presented and the final model is applied to assess the impact on rail path market shares of potential network changes. This is the second case to demonstrate how a rigorous travel behavior study can be carried out based on data from ADC systems.

Figure 1-2 Thesis Structure



Chapter Five summarizes the research findings including the characteristics of the ADC systems, rail trip OD matrix inference methods, and passenger path choice modeling; points out the limitations and difficulties in the current study; and proposes several future research directions.

Chapter Two: Automated Data Collection Systems: Potential and Reality

This chapter first describes the common characteristics of ADC systems including their advantage and disadvantages, and the reasons for these disadvantages. These characteristics are then illustrated by examining in detail two ADC systems recently implemented in the Chicago Transit Authority (CTA)—the Automated Fare Collection system (AFC) and the Automated Vehicle Location system (AVL). Thanks to their unique features, there is great potential for ADC systems to be used to support decision-making in transit agencies. However, it is often not straightforward to effectively utilize ADC system data. Several examples are given to demonstrate that there is a critical gap between what ADC data directly offers and what is needed in transportation planning and operations analysis. As a consequence, ADC data sources have never been utilized effectively in practice. A foundation including data processing and analysis with the support of advanced technologies such DBMS and GIS is required to bridge this gap so that the full value of these new data sources can be achieved.

2.1 General Characteristics of ADC Systems

Applying new technology to automated data collection has been an important trend in data collection in public transit agencies. ADC systems provide new data sources with distinct advantages over traditional data collection methods in many respects:

First, ADC systems offer a large, often full, spatial coverage of a public transportation system, in contrast to traditional methods, which usually focus only on key locations due to budget and staff limitations. ADC systems also have a large, often full, temporal span. They collect data 24 hours a day and 7 days a week, providing continuous data flow, in contrast to traditional methods, in which a "typical" day is chosen, normally a weekday with good weather and no special events, and often only its peak hours are considered. In today's complex society, special events, unusual weather, or other anomalous conditions occur every day in some part of the city. It becomes more and more difficult to find a "typical" day to represent the general reality. Thanks to their full spatial and temporal coverage, ADC systems offer the opportunity to study these atypical conditions, special events and weathers, both on weekdays and weekends.

Second, traditional methods are expensive in terms of both monetary budget and manpower. Therefore they can only be carried out at a very low frequency—once a year, or even once five years, for instance. In contrast, ADC systems incur a small marginal data collection cost once the systems are installed and so data is being updated constantly.

Third, ADC data are (by definition) automatically collected and digitally stored. With the support of an effective data processing and analysis framework, ADC data can be processed very fast. In traditional methods, by contrast, even after the field surveys are done, the survey forms are in paper form and it often takes months for the surveyors to input data and compile statistics before a clean data report can be produced.

Lastly, the accuracy of ADC data can be determined, avoiding all sorts of hard-to-detect survey errors that can appear in traditional methods.

ADC data also has inherent disadvantages derived from three main sources:

First, although some ADC systems were initially designed for data collection, many are not. For these, data collection is just an auxiliary function. For example, the Automatic Passenger Counter (APC) system is in the former category, directly aimed at collecting passenger count data on buses. In the latter category, however, the AVL system in CTA is part of an Automated Voice Annunciation System (AVAS), designed for announcing stop information on buses; and the AFC system is focused on monitoring fare collection and revenue counts. When these systems are looked at as data collection devices, if this was not their original design purpose, not surprisingly, many limitations arise. Some lack critical information. Some collect data for only a limited purpose such as revenue accounting, even though they could have been simultaneously collecting data for planning/operational usage at small incremental cost. The data collected are often intermittent, fragmented, and not in an easy-to-use format, which can make the raw data unintelligible and meaningless without intensive processing.

Second, different systems are implemented separately by different vendors. The resulting data structures and formats are often incompatible; the data are stored and managed in different database management systems; and ADC systems are implemented at different

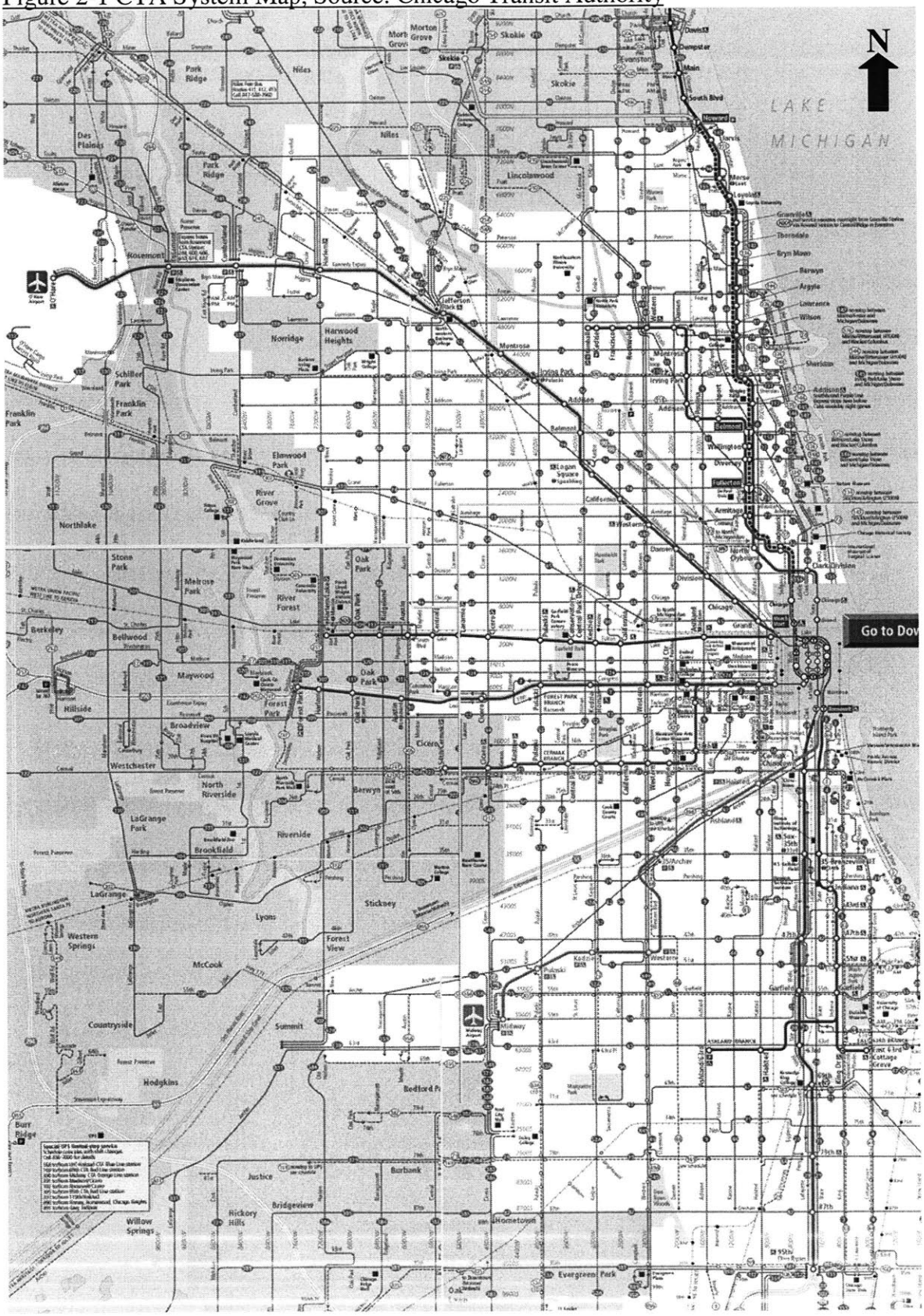
locations and in different time periods. All of this makes integration across systems very difficult and seriously limit the usage of ADC data.

Third, after ADC systems are installed and the vendor's technicians have left the agencies, the after-sale support can be a problem. Transit agencies generally lack professional staff that specialize in processing and analyzing data from ADC systems. Staff are required to be familiar with ADC systems, with technologies like database management, GIS, programming, etc., as well as with transportation analysis needs themselves. Because a framework for ADC system data manipulation incorporating analysis methodologies and techniques is not available, the utilization of ADC systems generally does not go beyond what was initially designed by the vendor.

2.2 The CTA ADC Systems

The Chicago Transit Authority (CTA), the nation's second largest transit system, delivers 1.5 million rides on an average weekday in Chicago and 38 suburbs. CTA's nearly 1,900 buses serve 1 million passenger trips daily serving nearly 12,200 bus stops. More than 1,100 CTA's rapid transit cars serve about a half-million passenger trips each day serving 143 stations. (See Figure 2-1)

Figure 2-1 CTA System Map, Source: Chicago Transit Authority



There are three ADC systems currently operating in the Chicago Transit Authority : the Automated Fare Collection system (AFC), the Automated Vehicle Location system (AVL) and the Automated Passenger Counting system (APC). This section does not intend to cover all the details of these systems; rather, it focuses on the AFC and AVL systems and only introduces the aspects that help illustrate the characteristics introduced in section 2.1 as well as those that are related to the processing and interpretation of ADC data.

2.2.1 The Automated Fare Collection system (AFC)

Entry Only

CTA's AFC system is an entry-only system, meaning passengers only use their farecards when entering a rail station or boarding a bus, and do not use the farecards when exiting the rail system or bus. Therefore, no information about a trip's destination is provided. Because the CTA pricing scheme is flat fare, irrespective of the travel distance, travelers are charged only when entering. Since the AFC system is designed only for fare collection and revenue accounting, it is not concerned with exiting from the system, and thus does not record destination information, even though this would be critically important for transit planning and operation.

Record Fields Description

A transaction record is generated each time a farecard is swiped on the AFC equipment.

The record includes the time, the location, the event and the unique serial number of that farecard. The time is stamped to the nearest second for each use. The location information is discussed later. The event field is reported to distinguish transaction types: these can be trip transactions such as entering stations or boarding buses, or non-trip transactions such as card purchases, or adding value. The transaction type is further refined to distinguish between initial use of a farecard and its subsequent use for a transfer. A unique serial number is assigned to each farecard, which allows tracking an individual rider's multiple trips over the lifetime of the farecard, which could be one trip, one day, one week or even one month.

Location Information

The location is recorded differently for train trips than for bus trips in the AFC system. For train trips, the location is recorded as the station entrance code. For bus trips, the location is coarsely represented by the bus route number (instead of the bus stop ID). Therefore, while the AFC records the exact boarding location for train trips, it doesn't record the exact boarding location for bus trips. For example, if a rider boards a bus on route 55 at Garfield/South Harper, the AFC system records the bus route number (55), but doesn't record the Garfield/South Harper bus stop. The coarseness of the location reporting for the bus trips causes difficulties in the utilization of the AFC data for some applications. One example is presented in section 2.3 to demonstrate the solution to the problem of how to determine the boarding location for bus trips more precisely.

AFC Coverage

Spatial and Temporal Coverage: The AFC system is installed on the entirety of both the CTA rail system and buses during all operating hours.

Usage Coverage: because the AFC farecard is not the only way passengers can pay for their trips, the AFC system does not provide farecard-type data for all passenger trips. The other payment methods include cash and paper transfer, both of which are categorized as non-serial encoded fare media since they don't have a serial number associate with the fare media. The farecard usage rates are different for rail trips and for bus trips. For rail trips, while cash payment by inserting coins is still possible in some turnstiles has become very rare, paper transfer from the previous bus trip segment still accounts for a sizable portion of rail trips. Both non-serial encoded fare media account for approximately 9%[1] of all rail trips, with farecard media covering the remaining 91%. Of course this coverage rate varies by the time of day and by station. The coverage rate on buses is lower with approximately 60% of bus trips paid by the farecard media in 2000.[2]

Pace Bus: In Chicago, there is another bus transit provider—Pace, which serves mainly suburban travelers. Pace serves 130,000 riders daily with 240 routes, 450 vanpools and many Dial-a-Ride programs. An AFC system is also installed on Pace buses, but was not fully integrated into the CTA data system.[3] However, travelers can use the same farecard to board CTA and Pace buses. In CTA's AFC data records, the field "Last Route" stores

[1] From CTA RFP: statistical validation of the farecard passenger flow model for federal NTD reporting
[2] Calculated as the average of 65% for commuter trips and 48% for non-commuter trips from CTA 2001' Travel Behavior and Attitude Survey, Pages 70, 83, 95.
[3] From spring, 2004 the data from PACE have been shared with CTA. The algorithm presented later in this thesis can be simplified due to this change.

the previous trip's information. If the previous trip is on a Pace bus, this field indicates the Pace bus route number.

2.2.2 The Automated Vehicle Location system (AVL)

The AVL system in CTA is part of the Automated Voice Annunciation System (AVAS), which is required by the Americans with Disabilities Act (ADA) to provide route and stop information to bus riders. The AVAS system provides bus location information along with the bus identification information, and the bus status information, and was implemented by Clever Devices, Inc, starting in October 2002.

The system records a number of events related to vehicle progress along a route. Each event has a time-stamp to the closest second, and the location information derived from the dead-reckoning enhanced GPS, as well as identification information such as bus route number and bus number, and status information such as whether the door is open or closed.

The location information is recorded as longitude and latitude coordinates and the route number, run pattern ID and the stop sequence number. Although the bus stop ID is the most meaningful location information for transit agencies, the AVL system does not report it directly. Later in section 2.3, I will present techniques needed to transform the route number, pattern ID and stop sequence number to the bus stop ID.

Coverage rate

About 1600 vehicles, or approximately 80% of the present fleet, have had AVAS hardware installed as of July 2004, and the system is expected to be fully operational by the end of 2004. All new vehicles purchased will have AVAS installed, in order to bring the percentage of equipped vehicles in the fleet to 100% as vehicles are retired and replaced. When the AVL data was collected for this study in January 2004, about 70% of the total bus fleet was equipped with AVL devices. However, not all devices were functioning throughout the time the bus was operating. Therefore the effective coverage rate (the ratio of the bus-hours covered by the AVL data over the total bus-hours operated) is much lower, roughly 40% as estimated in this study.

2.3 Potential and Obstacles for ADC Data Utilization

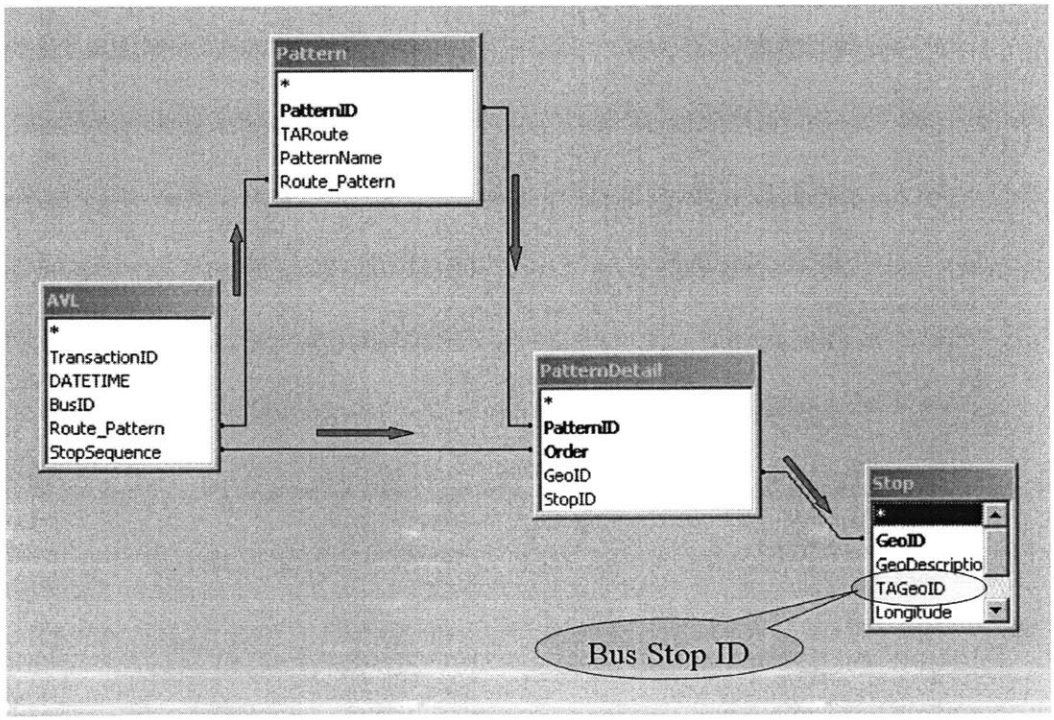
In introducing the AFC and AVL systems at the CTA, several problems were noted with regard to the use of data: the bus stop ID is not directly recorded by the AVL; while the bus route number is reported in AFC, the bus number and the specific boarding bus stop are not recorded. Examples are given below to show techniques that can be used to solve these problems and, more importantly, illustrate that while there is great potential for these new ADC data sources, there is a critical gap between what ADC data directly offers and what is desired by transportation planners and analysts. The complexity of the data manipulation tasks in the examples to be presented increases sequentially—from the processing of data from an individual ADC system, to the integration across multiple ADC systems, and to full scale AFC and AVL data processing.

2.3.1 Processing of data from an individual ADC system

AVL Data Processing—Identifying Bus Stop ID

The bus stop ID would be the most useful means of recording the bus location, but the AVL system does not record it directly in the raw data; instead, it reports bus route number, bus running pattern ID and stop sequence. Figure 2-2 is the entity-relationship diagram that illustrates the AVL data table and auxiliary tables involved (Pattern, Pattern Detail, and Stop) and their inter-relationships. The AVL table is updated everyday and the auxiliary tables are static in the long term. Relational database technology is a pre-requisite to the process of translating the route, pattern and stop sequence into a bus stop ID. The difficulty is not to understand the relationship once it is identified, but to figure out the relationship from the dozens of tables and their hundreds of possible inter-relationships.

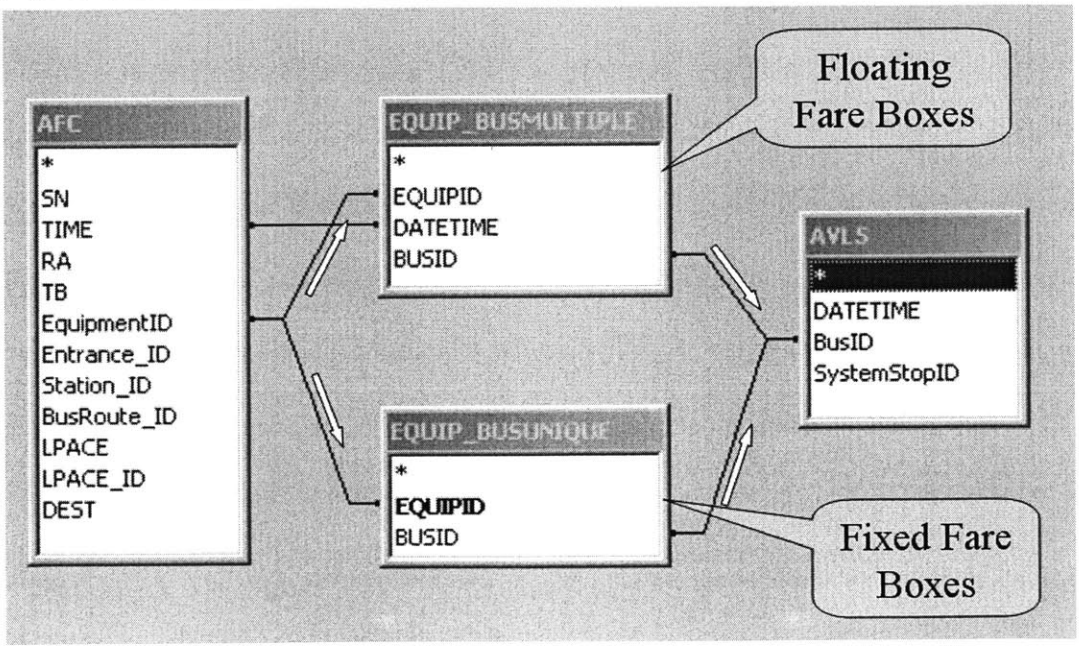
Figure 2-2: AVL Data Processing—Identifying Bus Stop ID



AFC Data Processing—Identifying the Bus ID

When the AFC system is used on buses, it records the bus route number but not the bus ID, which would be of greater use for fleet management. Figure 2-3 shows the technique that solves this problem by taking advantage of the relationship between the AFC fare box equipment ID and the bus ID. Most fare boxes are fixed on a specific bus, but some of them are floating among different buses at different times. In the relational database management system, two table joining operations are performed--one for the fixed fare boxes, for which a simple one-to-one relationship is sufficient; the other for the floating fare boxes, for which the time stamp is needed in addition to the equipment ID to complete the join.

Figure 2-3: AFC Data Processing—Identifying the Bus ID

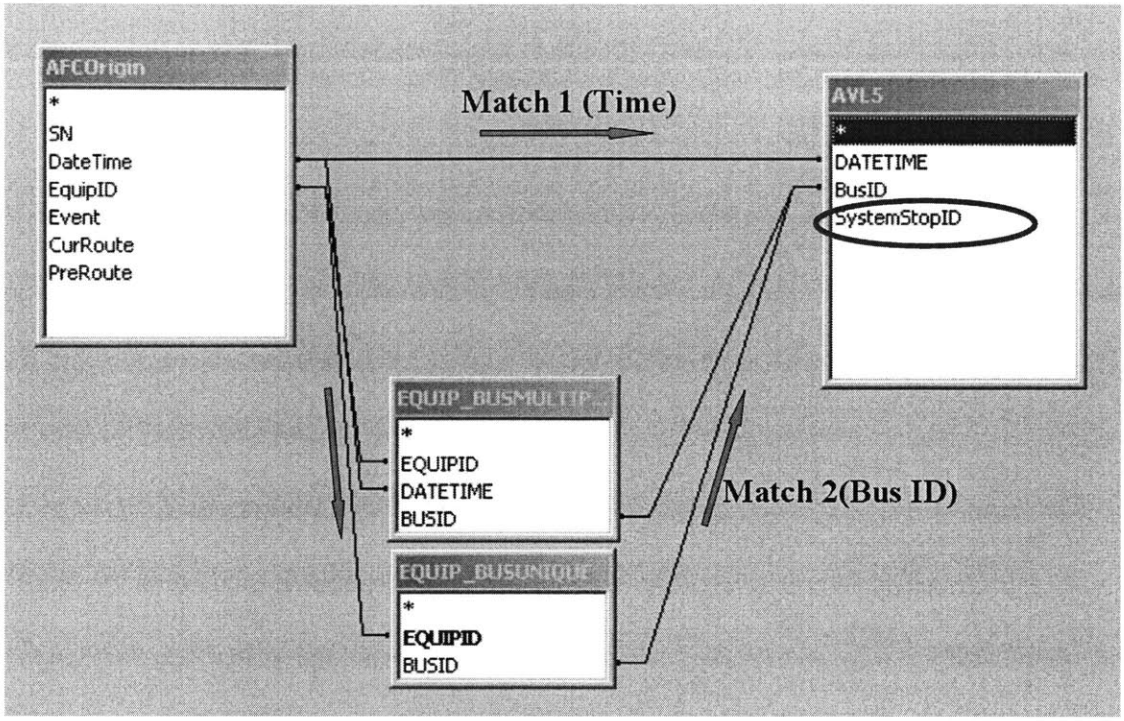


2.3.2 Integration across ADC systems to obtain the boarding bus stop

As argued in section 2.2 about the AFC system, the location information about the bus trip is very coarsely recorded. More specific boarding location—boarding bus stop—is important information to carry out many analyses about passenger bus trips such as bus trip OD analysis; however, the AFC system records the bus route number but not the bus stop ID. The combination of the AFC system and the AVL system offers a solution here. The AFC data records traveler trips while the AVL data records vehicle trips. In Section 2.3.1 I presented a method to obtain the exact bus stop ID for a vehicle trip from the AVL data. If the AVL vehicle location information can be matched against the AFC passenger trip information, this allows us to infer the boarding bus stop ID for specific trips.

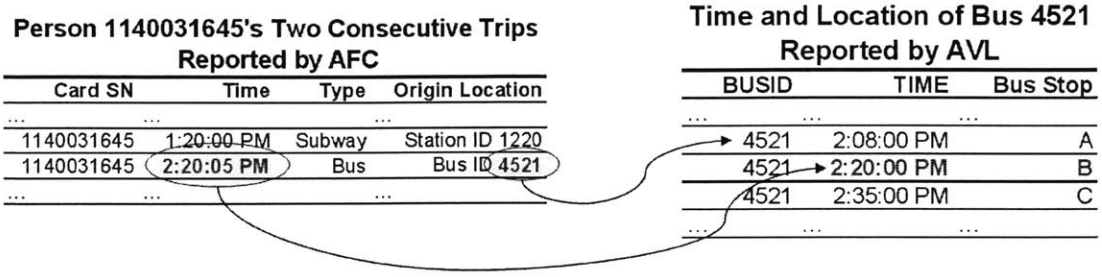
The combination of AFC and AVL is realized by two linkages: first, the time when a passenger swipes the fare card (data from AFC) matches the time when a bus arrives at a bus stop (data from AVL). The time stamp from the AFC should be shortly after the time stamp from the AVL. Secondly the bus ID from the AFC matches the bus ID from the AVL. If the two linkages can be established simultaneously, the bus stop that the vehicle is arriving at is then the bus stop where the passenger boards, as shown in Figure 2-4.

Figure 2-4: Combining AFC with AVL to get the boarding bus stop ID



For example, in the left table in Figure 2-5, the AFC records two trips of a person with the fare card 1140031645. The person boarded bus 4521 at 2:20:05pm. In the right AVL table, the three records for bus 4521 indicate that this bus passed stop A at 2:08:00pm, stop B at 2:20:00pm and stop C at 2:35:00pm. After matching the AFC table with the AVL table, it is safe to infer it is stop B where the person boarded at about 2:20pm. The two times 2:20:05pm from AFC and 2:20:00pm from AVL are not exactly the same. The story is that the AVL system reports the time 2:20:00pm when the bus arrived at stop B, and the passenger boarded the bus and then swiped the AFC card at 2:20:05pm. Bus stop B is then the boarding stop for this bus trip.

Figure 2-5: Example of AFC and AVL combination



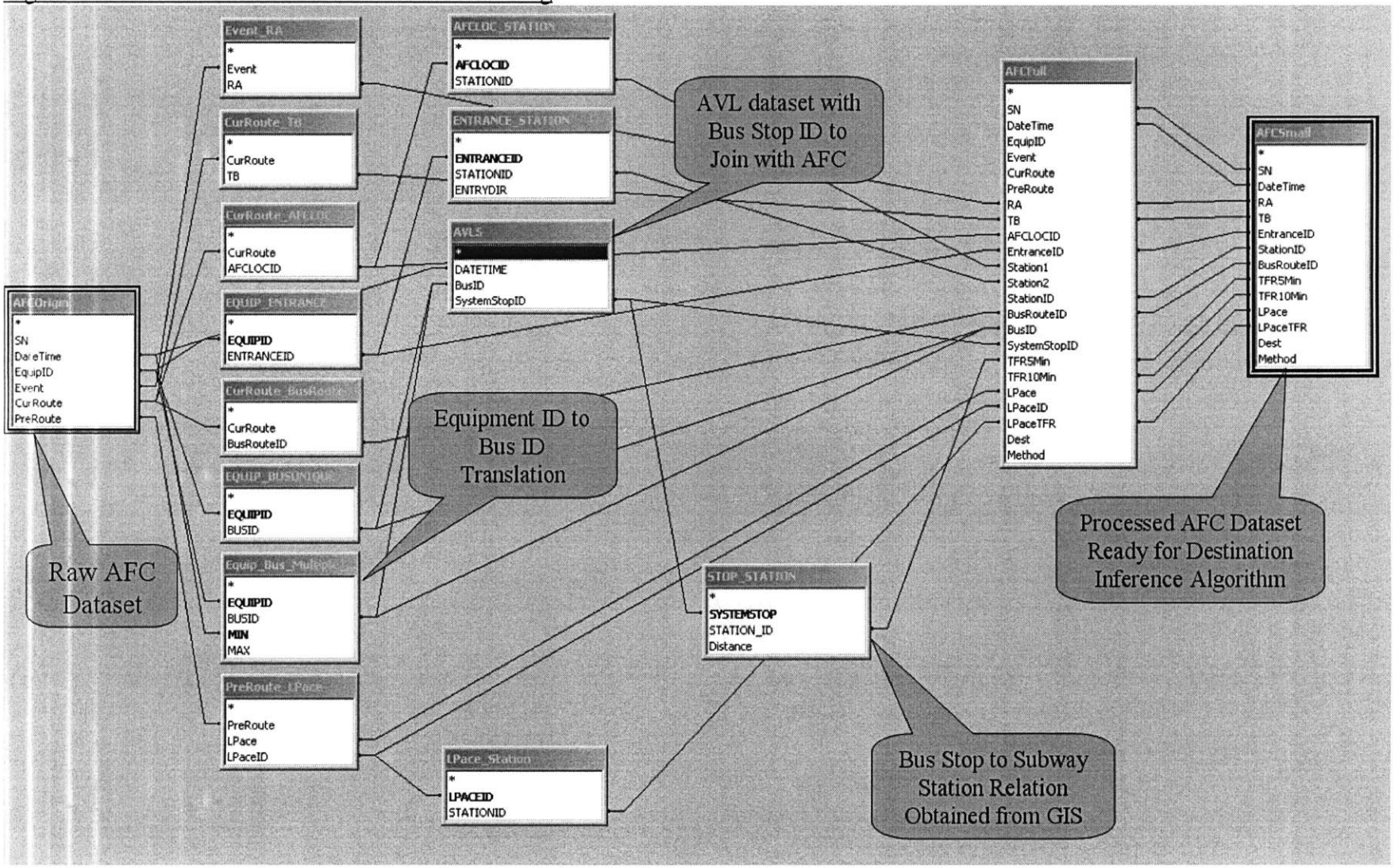
2.3.3 Full scale AFC and AVL data processing

With the above processing of the individual AFC and AVL data, and the combination of the AFC and AVL system data, we are now ready to perform the full scale AFC and AVL data manipulation, which transforms the AFC data from the raw format (indicated by the leftmost table in Figure 2-6) to data that public transit planners can make use of (indicated by the rightmost table in the same figure). The AFC and AVL tables are from the centralized AFC and AVL databases that are daily updated. The AVL table is processed to include a SystemStopID as shown in section 2.3.1. The Stop_Station relational table is derived from the GIS analysis of proximity between bus stops and rail stations. Other tables are reworked relational tables developed from ADC systems documents and specific data requests to the data department of CTA. The full scale AFC data processing combines the above examples of processing individual AFC data to get the bus ID, processing individual AVL data to get the stop ID, integrating AFC and AVL data to link the passenger trip information to the vehicle trip information to get passenger boarding bus stop ID, GIS analysis of the proximity, and other auxiliary steps such as table joining and data recoding.

There is a striking gap between what ADC systems offer and what transit agencies need. The dazzling joins, flows, tables, structures, relationships, etc. evidently says that it is not easy to bridge this gap. Meanwhile, a framework for ADC data manipulation and analysis methodologies and techniques is not readily available, and transit agencies generally lack staff familiar with the technologies such as DBMS, GIS, and programming, with ADC systems, and at the same time with transportation planning requirements. As a consequence, these data sources are not yet effectively utilized in practice.

It is conceptually a good idea to utilize ADC data but in order to make it practical in reality, a strong foundation of data processing and analysis with the support of advanced technologies such as DBMS and GIS is needed before the full value of the ADC data source can be exploited.

Figure 2-6: Full Scale AFC and AVL Data Processing



Chapter Three: CTA Rail OD Matrix Inference and Analysis

The rail passenger origin-destination (OD) matrix is of fundamental importance both for long-term rail infrastructure investment planning and for daily operations planning. Trip OD data are typically collected by specially designed passenger surveys. However, these surveys are usually expensive, subject to unknown bias and not easy to update frequently. The implementation of ADC systems provides a chance to automatically collect trip OD information at low marginal costs. The utilization of ADC data to obtain the trip OD entails different complexity for different ADC system configurations. In transit systems which operate with both entry-control and exit-control AFC systems (such as Bay Area Rapid Transit and Washington Metropolitan Area Transit Authority), the origin-destination matrix can be directly observed, however in systems such as Chicago Transit Authority and New York City Transit Authority, which have entry-only AFC systems, it is not as easy. This chapter examines the problem of how to infer the destination for a given origin for a passenger traveling in the rail network. This is the fundamental step in estimating a rail origin-destination matrix.

Section 3.1 proposes the idea of an improved destination inference algorithm that takes advantage of the pattern of individual transit rider's consecutive trip segments, and the integration of AFC and AVL systems. Section 3.2 describes the implementation of the

algorithm and the development of a software tool to automate the process. Section 3.3 applies the algorithm to the CTA rail system for a one-week period to infer the rail trip OD matrix and analyzes the CTA rail trip patterns from the estimated OD matrix.

3.1 Destination Inference

This study only deals with destination inference of rail trips (or trip segments[4]). Bus trip segments are utilized to help infer the rail trip segment destinations but the destinations of bus trip segments are not addressed.

The destination inference method takes advantage of the pattern of a person's consecutive transit trip segments. The method uses information on the next trip segment to infer the destination of a given trip segment. There are two different cases that are of interest in this study, rail-to-rail and rail-to-bus trip sequences. The rail-to-rail case has been discussed in Barry et al. (2001) and Rahbee (2002). This study distinguishes itself from the previous ones by examining the rail-to-bus case. Three strategies are developed to utilize the bus trip segment information for rail-to-bus sequences to different extents depending on whether AVL data are available and whether GIS technology is employed. In the CTA context, the inclusion of the rail-to-bus case by utilizing GIS technology and integrating the AVL data into the inference process increases successful rail trip destination inference by 10 percent.

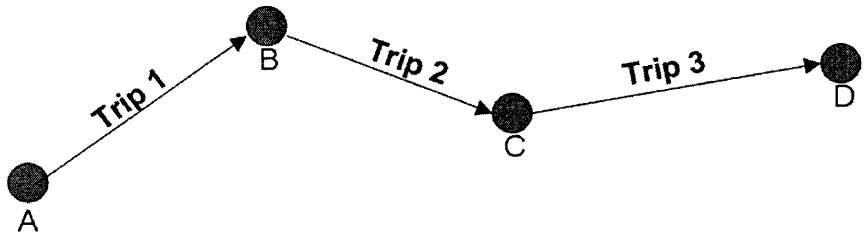
[4] A trip is composed of one trip segment if there is no transfer in the trip or more than one trip segments if there are transfers in the trip.

Another feature of the method is to explicitly examine the symmetry in passengers' trip segment chains. For trip chains whose symmetric pattern can be verified, the inference can be expanded from using only consecutive trip segments to using non-consecutive trip segments. This expansion enhances the success rate of the inference algorithm by a further 2.4 percent.

3.1.1 Inference Assumptions

The basic idea of the destination inference is that the destination of a trip segment is also likely to be the origin of the next trip segment. More loosely, a high percentage of riders stay at, or return to, the destination station of their previous trip segment to begin their next trip segment (Barry et al 2001).

Figure 3-1 A Person's Three Consecutive Trip Segments



For example, Figure 3-1 illustrates a person's three consecutive trip segments. Trip segment 1 goes from station A to station B, trip segment 2 from B to C, and trip segment 3 from C to D. The origins of these three trip segments are observed. The second trip segment's origin is station B. It is reasonable to infer that station B is also the destination

of the first trip segment. Similarly the origin of the third trip segment, station C, may be reasonably inferred also to be the destination of the second trip segment.

This premise is subject to the following assumptions[5]:

1. There is no private transportation mode trip segment (car, motorcycle, bicycle, etc) mixed in with the public transportation trip segments;
2. Passengers will not walk a long distance to board at a different rail station from the one where they previously alighted. The condition is equivalent to the distances between rail stations being longer than a typically acceptable walking distance, which in this research is set to be 1320 feet, five minutes' walk distance at a speed of three miles per hour,
3. Passengers end their last trip of the day at the station where they began their first trip of the day.

For example, Table 3-1 lists the four trip segments by the person with card 1109344130. Given first two assumptions, the destinations of the person's first three trip segments are inferred as Sedgwick, Armitage and LaSalle respectively. The last trip segment's destination is set to be the origin of the first trip segment of the day—Central Park, given the third assumption.

Table 3-1 Consecutive Train Trip Segments of Card 1109344130

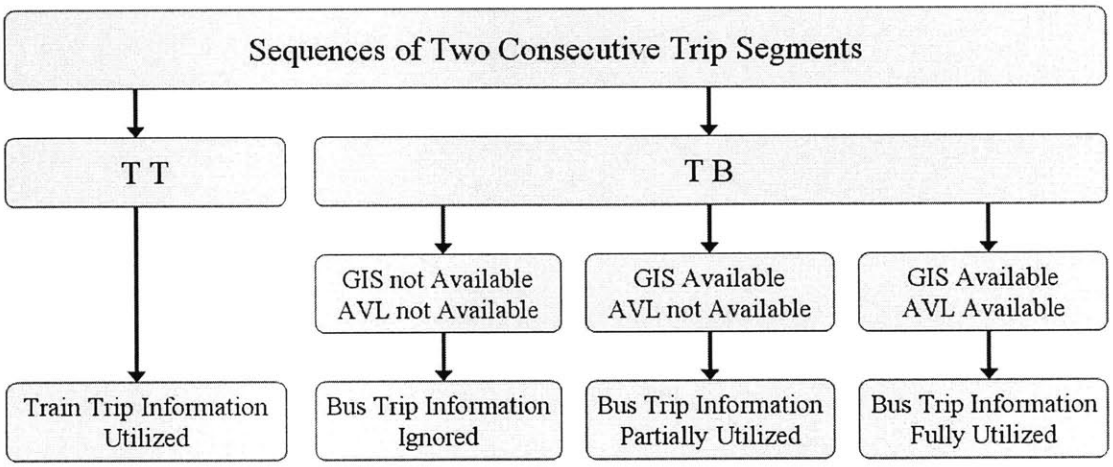
Transaction ID	Card SN	Time	Origin	Inferred Destination
377599	1109344130	34871	Central Park	Sedgwick
377600	1109344130	50486	Sedgwick	Armitage
377601	1109344130	77692	Armitage	LaSalle
377602	1109344130	78610	LaSalle	Central Park

[5] The validation of these three assumptions will be discussed later in Chapter five.

3.1.2 Two Consecutive Trip Segments

Since the algorithm uses the following trip segment's origin to infer the prior trip segment's destination, it is necessary to examine sequences of two consecutive trip segments. If "T" denotes a train trip segment, and "B" denotes a bus trip segment, two consecutive trip segments can be one of the four cases: TT, TB, BB, and BT. Because the study focuses on the rail trips, the BB and BT cases are not relevant. Therefore only train-to-train and train-to-bus cases are examined, as shown in Figure 3-2, and there are three situations in the TB case, determined by whether GIS technologies are employed and whether AVL data are available.

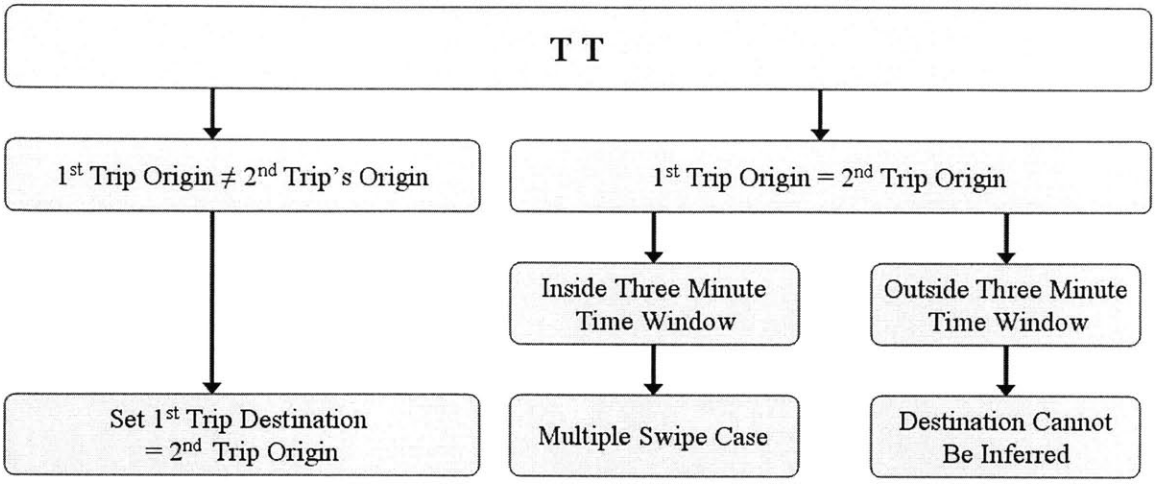
Figure 3-2: Cases of Two Consecutive Trip Segments



Rail-to-Rail Sequences

In the TT case, the second train segment's information can be directly utilized to infer the destination of the first segment. There are two situations as shown in Figure 3-3: first, if the origin of the second segment is different from the origin of the first segment, then the second segment's origin is assigned to be the first segment's destination; second, if the origins of the two segments are the same, then there are two scenarios.

Figure 3-3: Rail-to-Rail Sequences Scenarios



Scenario One: when the time stamps of both segments are within a short time window, the two segments are regarded as one travel group. This situation is labeled as "Multiple Swipe" indicating that a group of people are traveling together using a single farecard to pay for their trips. Three minutes is the time threshold used in the study. In this case all travelers in the group are assigned the same destination as the last traveler's destination given that this destination can be inferred. Scenario Two: when the time gap between both segments is beyond three minutes, the first segment's destination cannot be determined.

Rail-to-Bus Sequence

For rail-to-bus sequences, three strategies are developed for different system availabilities depending on whether AVL data are available and whether GIS technologies are employed. 1) When neither AVL data nor GIS technology are available, the bus trip segment information cannot be utilized. In this situation, the bus trips information is simply ignored as in Barry et al. (2001) and Rahbee (2002). 2) When GIS technology is

employed but no AVL data is available, the bus trip segment information can be partially utilized. 3) When both GIS technology and AVL data are available, the bus trip segment information can be fully utilized to enhance the destination inference algorithm.

No AVL Data but with GIS Technology

When GIS technology is employed but AVL data is not available, the bus trip information can be partially utilized. The utilization of the bus trip information requires that the spatial relationship between rail and bus network be analyzed. GIS provides tools for this type of spatial analysis. Buffers around rail stations are created as circles with rail stations as centers and 1320 feet as the radius. Intersecting the buffers with the bus routes can determine how many rail stations each bus route serves. There are three possibilities for how a bus route connects with the rail network, as shown in Figure 3-4.

In the first case when the bus route has no connection with the rail network, the destination of the train trip cannot be inferred. In the second case when the bus route has only one connection with the rail network, the connection station is the destination of the train trip segment. For example, bus route #17 goes from Westchester to Forest Park. It connects with the rail network only at Forest Park station on the Blue Line. A person with farecard 1095163393 boarded at Washington/Dearborn Station (Station ID=370) at 4:12PM and then boarded bus route #17 at 5:06PM as shown in Table 3-2. It can be inferred that the destination of the rail trip segment is Forest Park Station, the only connection between bus route 17 and the rail network.

Figure 3-4 Scenarios in the Train-to-Bus Sequence without AVL Data

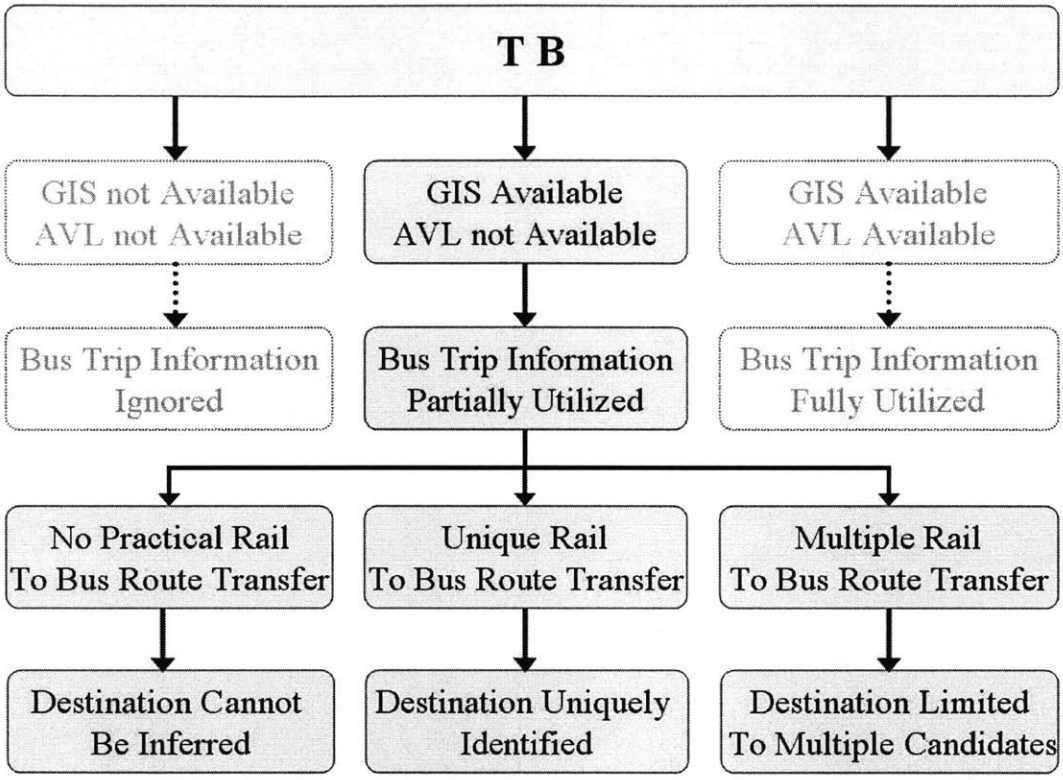


Table 3-2 Consecutive AFC Records for a Unique Rail to Bus Transfer Example

Consecutive AFC Records for Farecard 1095163393 on Jan. 14th, 2004

Card SN	Time	Type	Origin Location
...	...	...	
1095163393	4:12:50 PM	Rail	Station ID 370
1095163393	5:06:51 PM	Bus	Bus Route 17
...	...	...	

In the third case when the bus route intersects the rail network at more than one location, as shown in Figure 3-5, the destination can be limited to one of these candidate stations but cannot be uniquely identified. For example, a person with farecard 1028969715 boarded at Harlem/Lake Station (ID=20) at 4:54PM and then boarded bus route #55 at 5:54PM as shown in Table 3-3. This person could transfer from the rail network to bus route #55 at any of three connection points—Garfield Station on the Green Line, Garfield

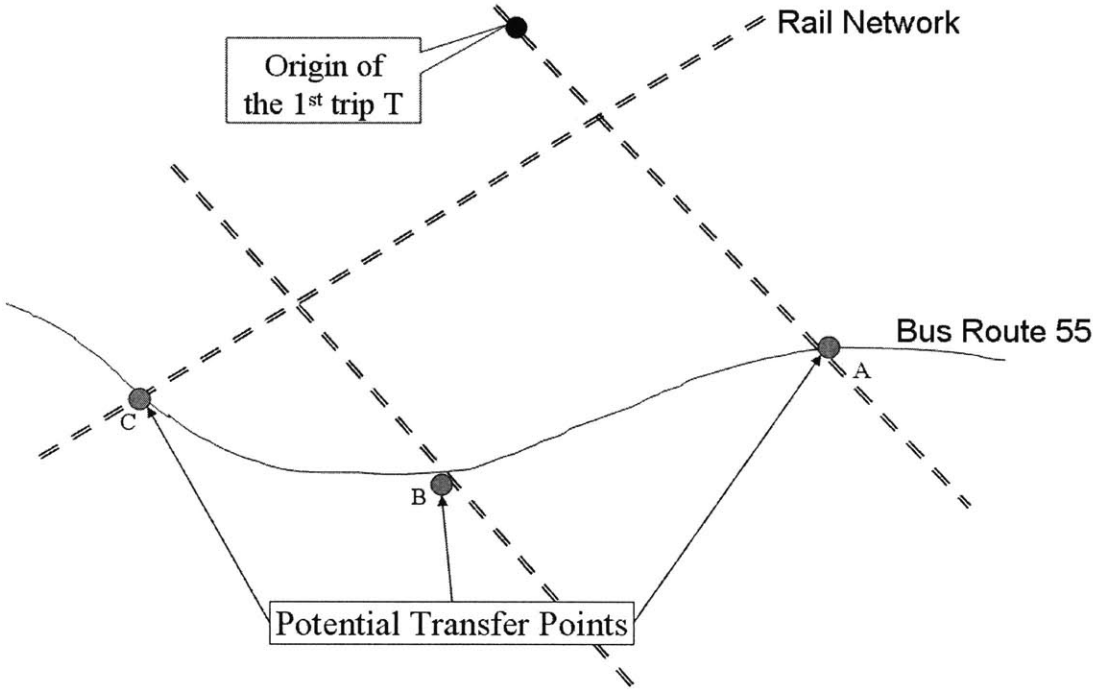
Station on the Red Line or Midway Station on the Orange line. It cannot be determined at which station the person actually transferred.

Table 3-3 Consecutive AFC Records for Multiple Rail to Bus Transfer Options Example

Consecutive AFC Records for Farecard 1028969715 on Jan. 13th, 2004

Card SN	Time	Type	Origin Location
…	…	…	
1028969715	4:54:41 PM	Rail	Station ID 20
1028969715	5:54:27 PM	Bus	Bus Route 55
…	…	…	

Figure 3-5: Multiple Rail to Bus Transfer Possibilities



With Both AVL Data and GIS Technology

When both GIS technology and the AVL data are available, the boarding bus stop of the bus trip segment can be identified through the integration of AFC and AVL data as described in Chapter Two; and the alighting station of the rail trip segment can be found using GIS to analyze the proximity between rail stations and bus stops in order to determine the station close to a bus stop, as illustrated in Figure 3-6. For bus stops with a

rail station within walking distance, the rail station is assigned as the train segment's destination. For bus stops where there is no rail station close by, the destination of the train segment cannot be inferred. Both cases are shown in Figure 3-7.

Figure 3-6: Destination Inference for TB Case with AVL data available

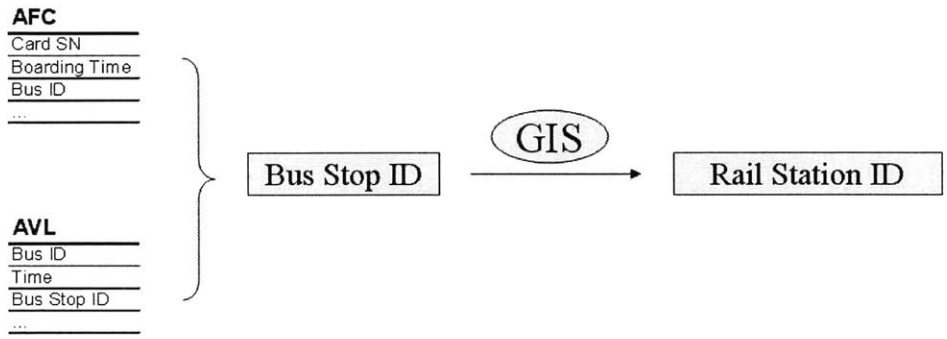
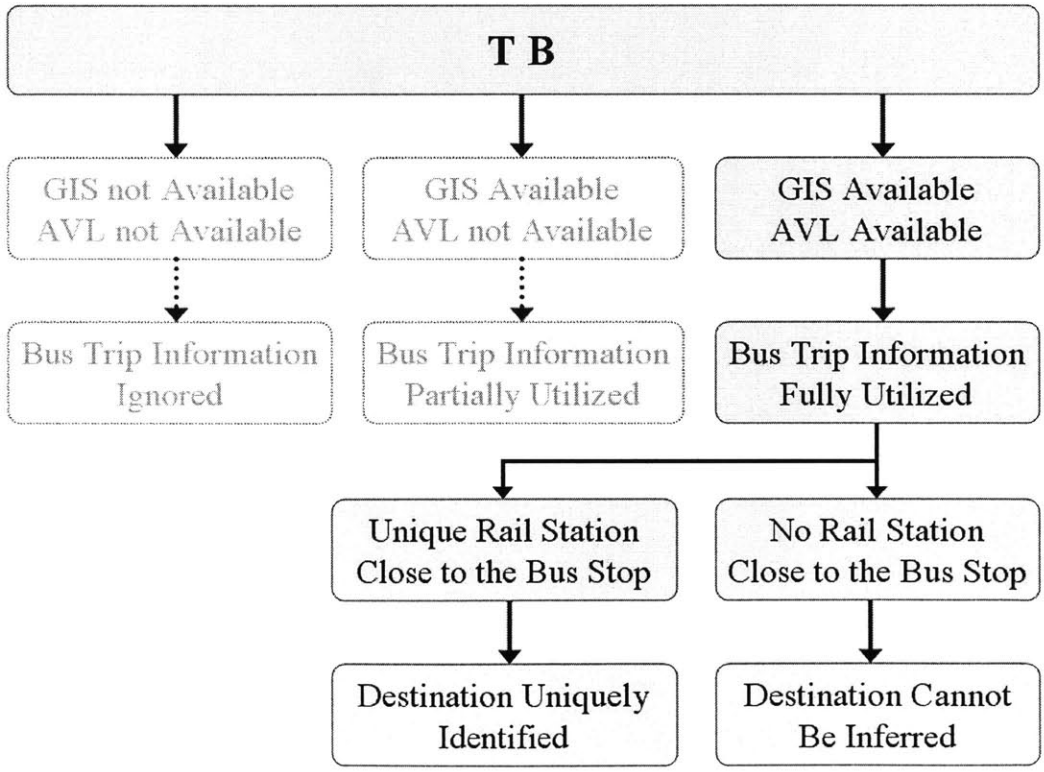
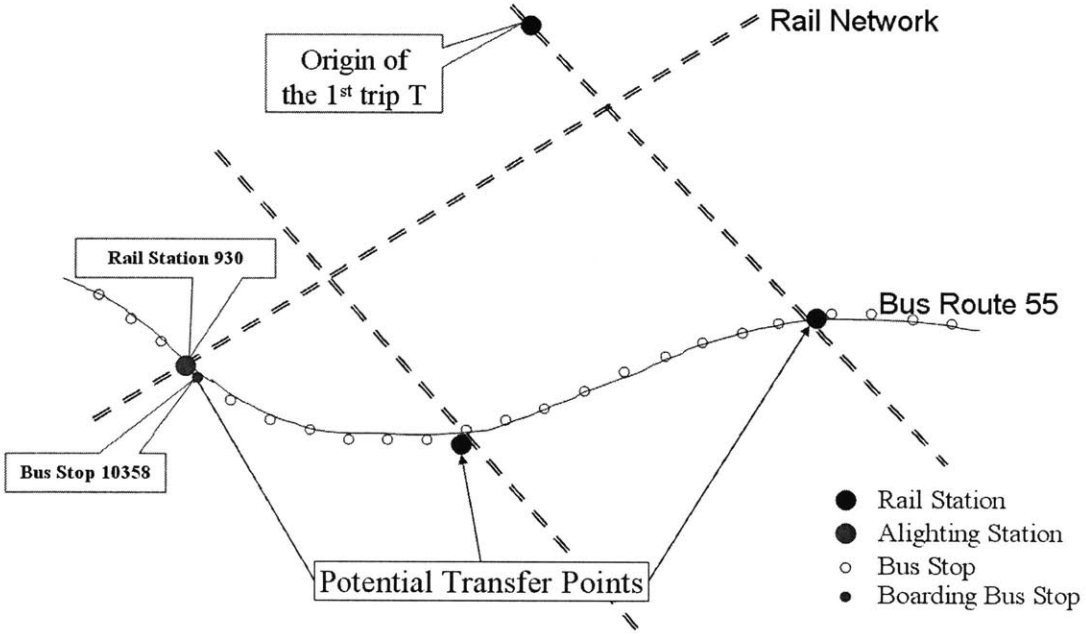


Figure 3-7 Scenarios in the TB Transfer Case with AVL Data



Continue the same example used in previous section. The person boarded bus route #55 at 5:54PM. By combining the AFC and AVL data, the exact boarding bus stop can be determined as bus stop 10358. And rail station 930 (Midway on the Orange Line) is the only rail station close to this bus stop. Therefore station 930 is assigned as the destination of the train segment, as shown in Figure 3-8.

Figure 3-8: Proximity between Bus Stop 10358 and Rail Station 390



3.1.3 Symmetrical Trip Chain Pattern Utilization

In this study the daily trip segment chains are utilized to improve the destination inference algorithm. Many trip segment chains exhibit symmetry in the whole, or in part of, the chain. For example, in the chain "TBBT" or any chains that include "TBBT" as a portion, the algorithm can check whether the bus route numbers of both bus trip segments are the same to verify the symmetry. When the symmetry is confirmed, the later T

segment can be used to infer the destination of the first T segment so that the inference possibility is expanded from using only consecutive trip segments to non-consecutive trip segments with symmetrical patterns. In the CTA context, this expansion allows 2.4% more rail trip destinations to be inferred.

3.2 Algorithm Implementation

3.2.1 Implementation Difficulties and the Development of "Rail OD Inference Tool" Software

The implementation of the algorithm is difficult and cumbersome, which requires the implementer be familiar with ADC systems and technologies such as DBMS, GIS and programming, as well as transportation planning. The processing of the AFC and AVL data and the application of the destination inference used to take the author about one month to complete a one-day rail OD matrix inference before the algorithm was developed. The complexity and time for the inference process may be acceptable for a scholar carrying out scientific research, but it is not realistic for transit agencies to implement in practice.

The "Rail OD Inference Tool" software is developed to facilitate this process. This tool allows a layperson without knowledge of the details of the ADC systems or in-depth understanding of database and GIS to be able to complete the destination inference process in a reasonably short time. This software tool takes the raw AFC and AVL data and several accessory tables as input, estimates the rail OD matrix automatically and

outputs it. Users only need to specify the input files and several parameters. As a performance measure, the software completes the inference of a one-week trip OD matrix for the CTA rail network in about one hour. The efficiency improvement results from two factors: 1) the automation of sequences of manual steps in the database processing such as table joining and queries; 2) and the speeding up of the algorithm by using customized, lower-level programming. The software is developed in the C/C++ programming language. The fundamental structure of the algorithm is described below.

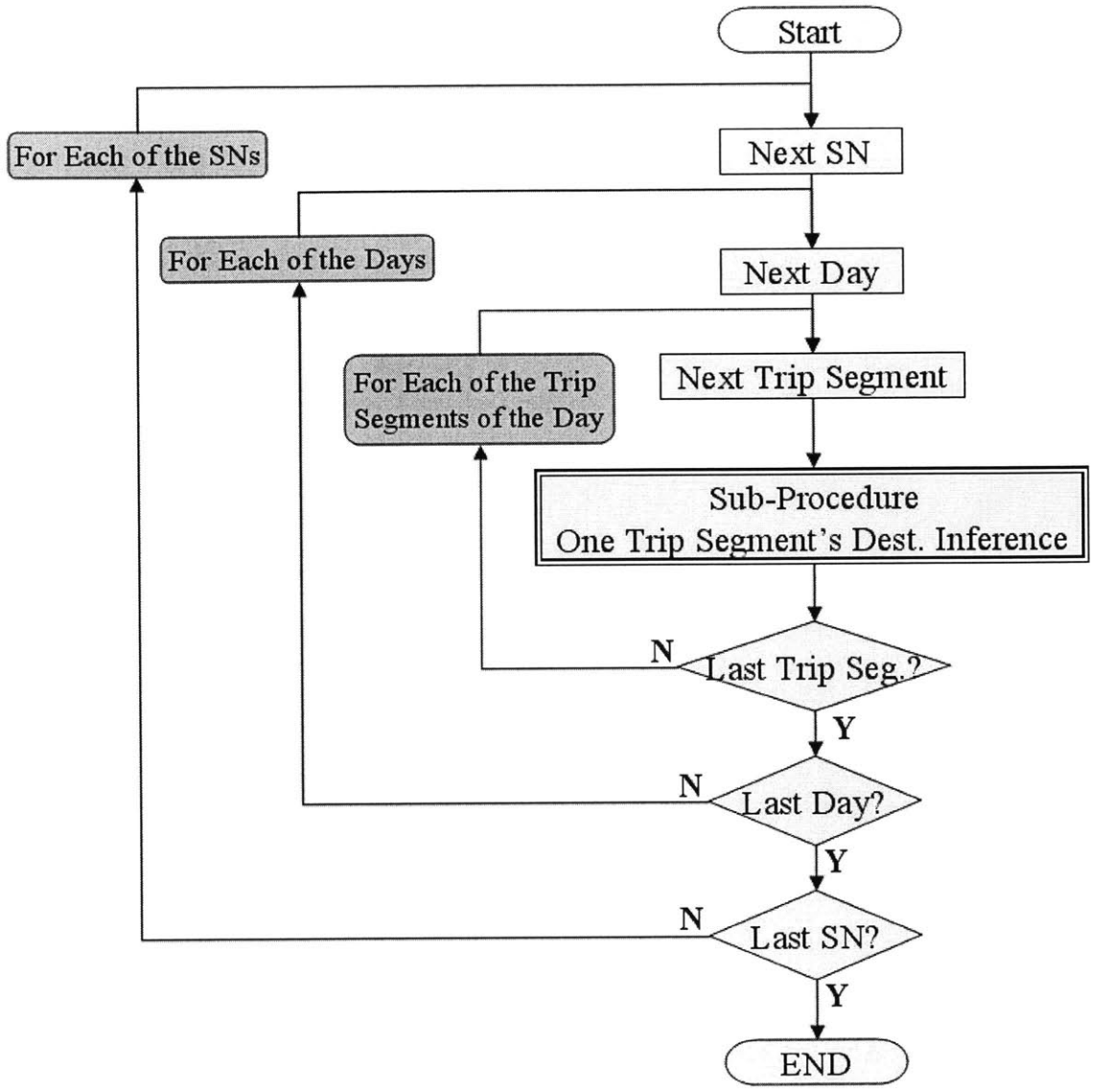
3.2.2 Inference Algorithm Structure

The algorithm is composed of three nested loops, a sub-procedure for one trip segment's destination inference, and separate procedures for train-to-train, train-to-bus, last-trip-of-the-day and previous-Pace-bus trip sequences.

Nested Loop Structure

The algorithm starts by sorting the whole AFC data table using the serial number (SN) of the farecard as the primary key and the transaction date and time as the secondary key so that the trip segments of the same farecard are grouped together in chronological order. As indicated in Figure 3-9, there are three nested loops in the algorithm. The outer loop cycles through each of the farecards, the middle loop through each of the days and the inner loop through each of the trip segments of the farecard identified in the outer loop on the day pointed at in the middle loop. For each rail trip segment, a sub-procedure is carried out to estimate the destination of the trip segment.

Figure 3-9: Nested Loop Structure of the Algorithm

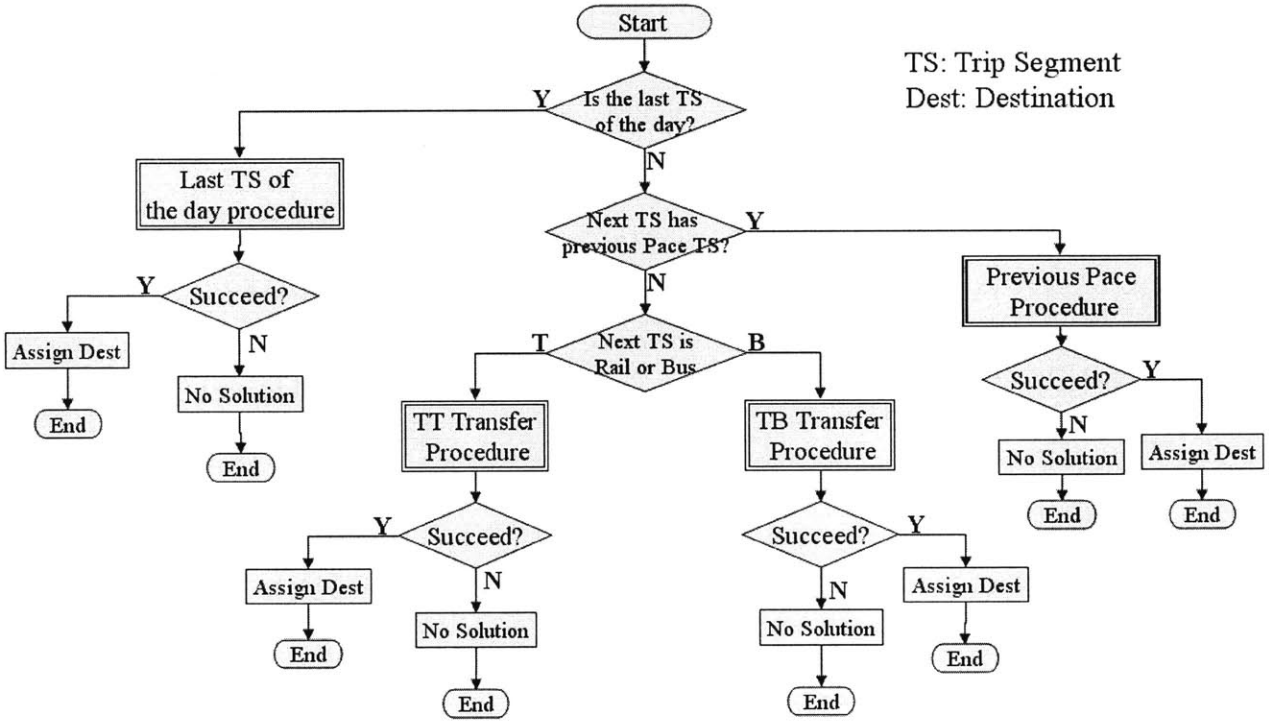


Sub-Procedure for One Trip Segment's Destination Inference

Figure 3-10 shows the overall sub-procedure structure. Suppose the rail trip segment currently under examination is trip segment *j* by person *p* on day *d*, and its immediate following trip segment is *k*. The algorithm determines whether trip segment *k* is on rail or on bus and moves onto "TT transfer procedure" or "TB transfer procedure" accordingly

except for two special situations: 1) when trip segment *j* is the last trip segment on day *d*, because there is no next trip segment for the last trip segment, a special "Last Trip Segment of the Day Procedure" is called; 2) when a Pace bus trip segment intervenes before trip segment *k*, the algorithm diverts to the "Previous Pace Procedure" to handle this special situation. Each of these four procedures is described as below.

Figure 3-10: Sub-Procedure for One Trip Segment's Destination Inference

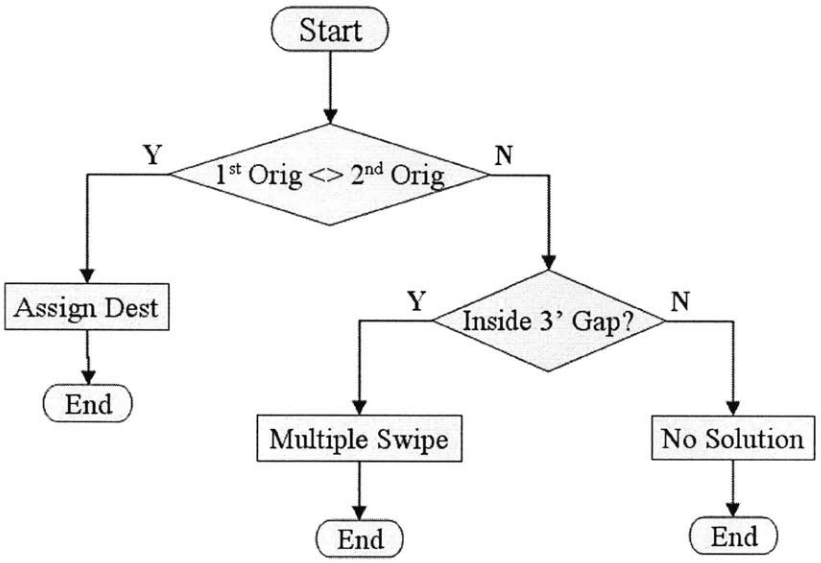


TT Procedure

The inference concept for the train-to-train case was introduced in section 3.1.3, and is implemented as two levels of bifurcations in the TT procedure, as shown in Figure 3-11. The only complexity lies in how to deal with the Multiple Swipe case. A "stack" is used to model the Multiple Swipe case—when consecutive segments are identified to be of the

same travel group, they are pushed into the stack one by one without knowing their destinations until it comes to the last transaction of the travel group—if the last transaction's destination can be inferred, then all trip segments in the stack are popped and assigned the same destination; if it cannot be inferred, then neither can any of the other trip segments in the stack.

Figure 3-11 TT Transfer Procedure

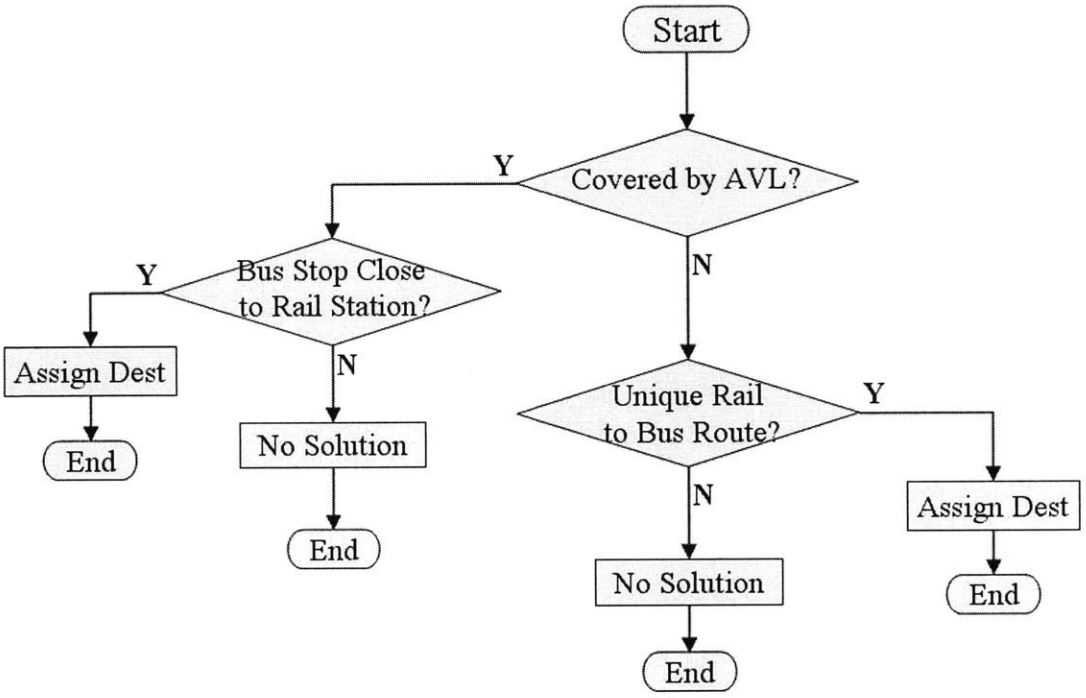


TB Procedure

In the TB procedure, the algorithm first checks whether the bus trip segment is made on an AVL-equipped bus and proceeds accordingly with different inference strategies. For the AVL-equipped bus trips, the boarding bus stop is obtained and the closest rail station to the bus stop is identified. If the bus stop and rail station are within walking distance, this station is assigned as the destination of the train trip segment. If the bus stop and rail station separation is beyond walking distance, then the destination of the train trip

segment cannot be inferred. Since the current CTA AVL system does not cover all buses, the algorithm turns to the secondary method for non-AVL-equipped bus trips. The rail to bus route transfer possibility is examined. If there is only one transfer station from the rail network to the bus route, this station is then assigned to be the destination of the train trip segment; otherwise, no destination can be inferred for the train trip segment. (See Figure 3-12)

Figure 3-12 TB Procedure



Last Trip of the Day and Pace Bus

When trip segment *j* is the last trip segment on day *d*, the third assumption that "passengers end their last trip of the day at the station where they began their first trip of the day" is utilized. The "Last Trip Segment of the Day Procedure" checks the first trip segment on day *d*. If the origin of the last trip segment is different from the origin of the first trip segment, the first trip segment's origin is assigned to be the last trip segment's

destination. The trip count by hour-of-day indicates that there are the fewest trips occurring between 2 and 3am so the day boundary is defined at 3:00am instead of midnight to better reflect actual behavior.

When there is a Pace bus trip segment before trip segment *k*, the "Previous Pace Procedure" is carried out. Pace buses are treated similarly to CTA buses with no AVL system installed. The transfer possibilities from the rail network to Pace bus routes are examined. When there is one and only one transfer station from the rail network to a certain Pace bus route, the transfer station is inferred to be the destination of the rail trip segment.

3.3 Algorithm Application for CTA Rail Trip OD Analysis

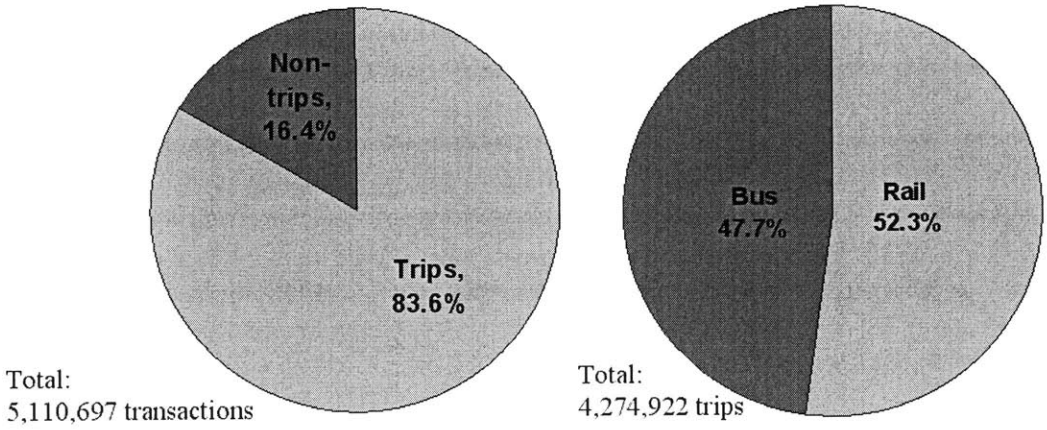
The algorithm is applied to the CTA rail system for a one-week period from Jan.11-17, 2004. The CTA AFC and the AVL data were input to the "Rail OD Inference Tool" and the corresponding rail trip OD matrix produced. The rail trip OD pattern and the trip segment chain pattern are analyzed based on the OD matrix produced by the algorithm.

3.3.1 Basic Statistics

There are a total 5,948,454 transactions recorded by the AFC system in the one week period analyzed including 4,974,422 trip transactions (83.6%) and 974,032 non-trip transactions (16.4%). The trip transaction mode split is 52.3% by rail (2,602,819 trip segments) and 47.7% by bus (2,371,603 trip segments), as shown in Figure 3-13. The

5,948,454 transactions are generated by 1,076,867 distinct farecards giving an average usage rate of 5.5 transactions per card in this week.

Figure 3-13: Trips vs. Non-Trip Transactions and Public Transit Mode Split



Out of the 2,602,819 rail trip segments, the algorithm successfully infers the destinations of 1,705,954, or 65.5%. The 35% loss is explained in the following examination of daily trip chain patterns. The following analyses are based on the trips that have a destination inferred by the algorithm. Table 3-4 summarizes the contribution of different inference methods. The inclusion of the train-to-bus cases accounts for 11% of the destination inference; the utilization of the symmetric trip chain pattern accounts for 2.4%; and the multiple swipe cases account for 1.4%. A total of about 15% of the destinations are provided by the improvement to the inference algorithm.

Table 3-4: Inference Methods Contribution

	Method	Percentage
Train-to-Train Case	Next Train Trip	51.0%
	Last Train of the Day	33.3%
Train-to-Bus Case	Unique Rail to Bus Route	5.1%
	AVL Selection of Bus Stop	5.9%
Special Cases	Trip Chain Symmetry	2.4%
	Multiple Swipe	1.4%
	Unique Rail to Pace Bus	0.9%
	Total	100.0%

3.3.2 Regional Trip Flow Pattern

In order to understand the regional trip flow pattern at a macro level, stations are grouped into four regions—the Loop, the North, the South, and the West. The Loop group includes 16 stations physically inside the elevated line circle plus LaSalle Station on the Blue Line that is very close to the Loop. The North group includes 45 stations on the Yellow, Purple, Brown lines, and the north portion of Red Line. The South group includes 22 stations on the Green Line south portion and the Red Line south portion. The West group includes 59 stations on the Blue and Orange Lines and the Green Line west portion, including O'Hare and Midway.

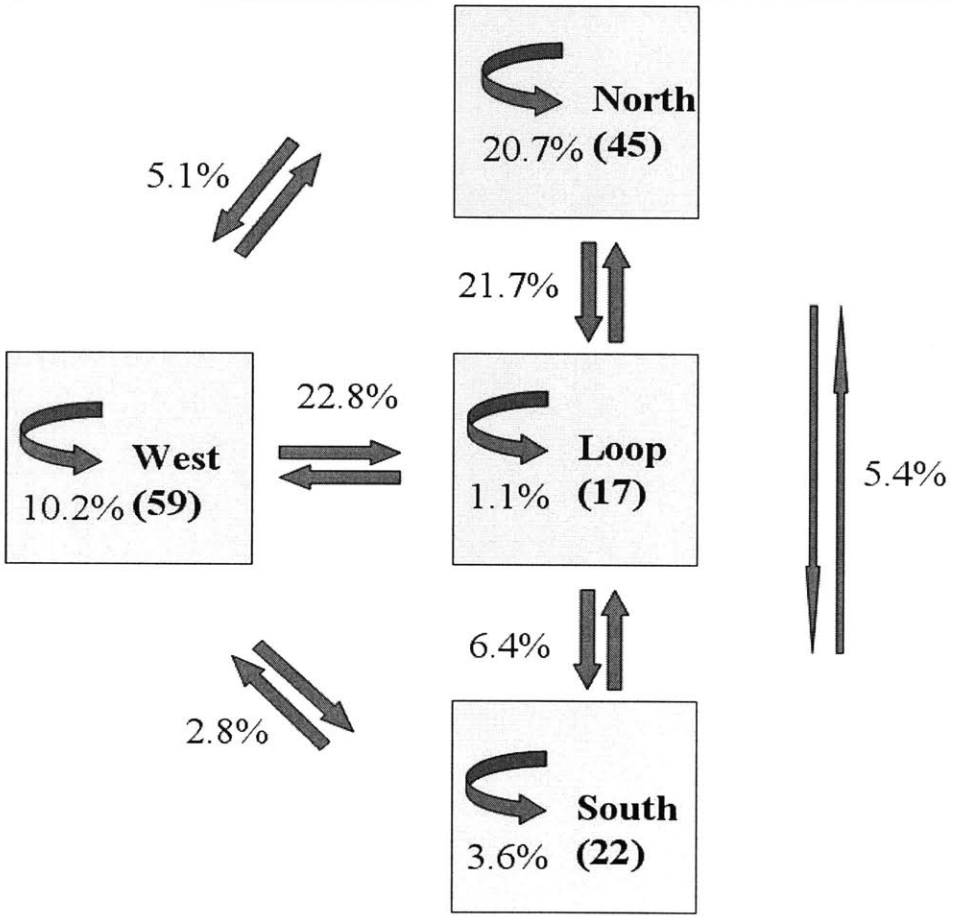
Table 3-5 shows the percentages of trips internal to each of the four regions. The North region has the highest number of intra-region trips with 20.7% of total trips. This is reasonable since the North region actually includes a large portion of the Chicago CBD. The Loop region has the lowest number of intra-regions trips, since the region is so small and dense that walking becomes the most convenient mode for these trips.

Table 3-5: OD Flow by Region

Orig	Dest	Trips	Percentage
North	North	352443	20.7%
West	West	173196	10.2%
South	South	62056	3.6%
Loop	Loop	19263	1.1%

Figure 3-14 also illustrates the trip flows between regions. The trip flows between the three peripheral regions are small, 5.1% between the North and West, 5.4% between the North and South, and 2.8% between the West and South. The largest numbers of inter-regional trips are between the West and the Loop, and between the North and the Loop.

Figure 3-14: OD Flow Inside or Between Loop, North, South and West Regions



3.3.3 Transit Trip Segment Chain Pattern

A transit trip segment chain on each farecard is the sequence of transit trip segments made by the farecard during a time period such as a day or a week. The typical daily trip segment chains look like "TT", "TBT", and "BTTB", while "0.T.T.T.T.T.0", "0.0.TB.0.0.0.0", "B.0.B.0.0.0.0" are examples of weekly trip chains[6] in which, a "." denotes the boundary between days and a "0" indicates that there are no trips on that day. Examination of trip chains helps us understand travelers' behavior of chaining trip segments in the day travel plan and the changes of travel pattern from day to day. One byproduct of the "rail trip OD inference tool" software is to generate daily and weekly trip segment chains. The following analysis is based on the trip chains observed in the one week period from Jan. 11-17, 2004 in the CTA system. Note that in this section the chain is simply defined at the modal level and does not have a spatial aspect.

Daily Trip Chain Pattern

On Monday, January 12, 2004, there were 359833 trip chains covering 1915 patterns, or on average of 188 chains per pattern. The top five most frequent patterns are "TT", "T", "B", "BB" and "TB", consisting of 69.3% trip chains as shown in Table 3-6.

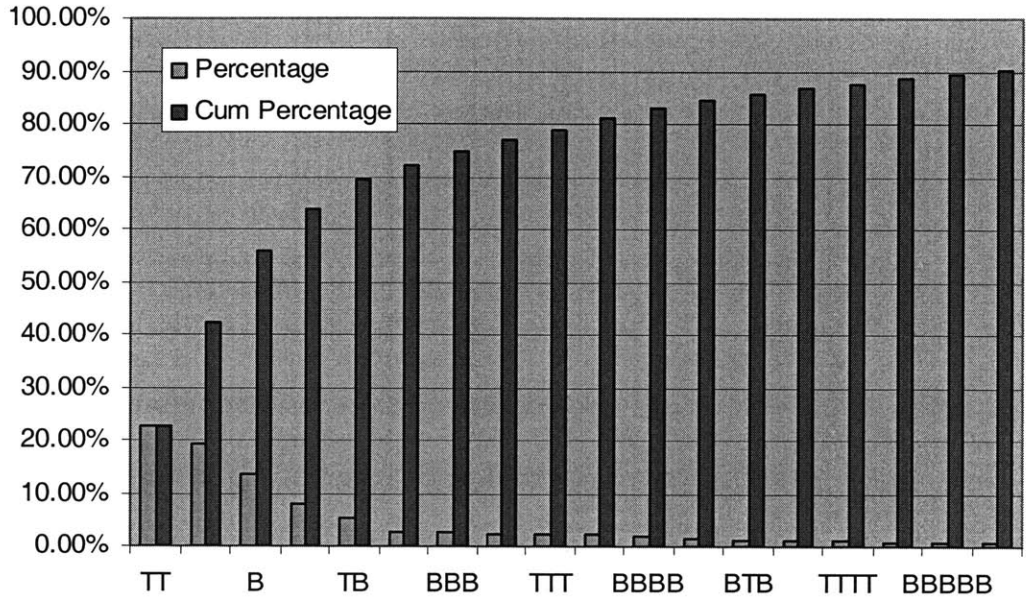
Table 3-6: The Top Five Most Frequent Daily Trip Patterns

Trip Chain Pattern	Percentage	Cumulative Percentage
TT	22.8%	22.8%
T	19.4%	42.2%
B	13.7%	55.9%
BB	7.9%	63.8%
TB	5.5%	69.3%

[6] Sunday is the first day in the weekly trip chain.

The top 18 most frequent patterns cover 90% of the chains as shown in Figure 3-15. The remaining 10% of chains are scattered among 1897 different patterns. Thus passenger travel patterns are highly diversified.

Figure 3-15: The Top 18 Most Common Daily Trip Patterns



The examination of the daily trip chain pattern helps us understand the performance of the destination inference algorithm: "T", a single rail trip in a day, for which no destination can be inferred, consists of 19.4% of all trip chains. This accounts for the biggest portion of the trip destination inference loss. "TB" or "BTB" are examples in which the enhanced algorithm outperforms the prior studies by examining the rail-to-bus cases. "TBBT" is an example of the improvement of the algorithm by utilizing the symmetrical trip chain patterns.

If the daily trip segment chains are compared between weekdays and weekends, weekdays are very stable, but, as expected, weekends are quite different. Table 3-7 compares the number of chains and number of patterns between weekdays and weekends,

and indicates that the trip chain patterns are less concentrated on weekends than on weekdays.

Table 3-7: Trip Chain Concentration by Day-of-Week

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
no. of chains	359833	365509	364470	363695	358351	167350	118035
no. of patterns	1915	1900	1867	1991	2100	1356	1115
pattern oncentration (chains/pattern)	**187.9**	**192.4**	**195.2**	**182.7**	**170.6**	**123.4**	**105.9**

The most common daily trip chain pattern also changes from TT on weekdays to T on weekends as shown in Table 3-8.

Table 3-8 Most Common Daily Chain Patterns by Day-of-Week

Mon		Tue		Wed		Thu		Fri		Sat		Sun	
TT	22.8%	TT	23.2%	TT	23.1%	TT	22.4%	TT	20.8%	T	21.9%	T	23.2%
T	19.4%	T	17.9%	T	17.7%	T	18.1%	T	19.8%	TT	15.9%	TT	14.9%
B	13.7%	B	14.1%	B	14.3%	B	14.5%	B	14.8%	B	12.0%	B	12.1%
BB	7.9%	BB	8.1%	BB	8.2%	BB	7.9%	BB	7.7%	BB	9.2%	BB	9.8%
TB	5.5%	TB	5.1%	TB	5.0%	TB	4.9%	TB	5.1%	TB	5.9%	TB	6.5%

Weekly Trip Chain Pattern

There were 1,076,867 distinct farecards used in the week analyzed with 176,546 different weekly trip chain patterns. So each pattern has, on average, 6.1 instances. The top five patterns are "0.0.0.0.0.0.0" (the card was only used for non-trip transactions), "0.0.0.0.0.T.0", "0.T.0.0.0.0.0", "0.0.0.0.0.B.0" and "0.0.0.0.B.0.0", consisting of 21.9% of weekly chains. The top 24 patterns cover only half of total weekly trip chains (see Table 3-9) with the remaining trip chains very sparsely distributed among 176,522 patterns. Thus weekly trip patterns are highly diverse. In particular, the degree of repetition of the daily chain in the weekly chain is far less than expected, i.e., many

passengers change their daily trip chain pattern during the week. In Table 3-9, only one pattern with the traditional repetition of the daily trip chain in the weekly trip chain is "0.TT.TT.TT.TT.TT.0", where a traveler makes two train trips every day on the weekdays.

Table 3-9: Most Common Weekly Trip Chain Patterns

Week Trip Chain Pattern	Counts	Percentage	Cumulative Percentage
0. 0. 0. 0. 0. 0. 0	109539	10.2%	10.2%
0. 0. 0. 0. 0. T. 0	32775	3.0%	13.2%
0. T. 0. 0. 0. 0. 0	32236	3.0%	16.2%
0. 0. 0. 0. 0. B. 0	31264	2.9%	19.1%
0. 0. 0. 0. B. 0. 0	30282	2.8%	21.9%
0. 0. 0. B. 0. 0. 0	29414	2.7%	24.7%
0. 0. B. 0. 0. 0. 0	28788	2.7%	27.3%
0. B. 0. 0. 0. 0. 0	27652	2.6%	29.9%
0. 0. T. 0. 0. 0. 0	26984	2.5%	32.4%
0. 0. 0. 0. T. 0. 0	26396	2.5%	34.9%
0. 0. 0. T. 0. 0. 0	25581	2.4%	37.2%
0. 0. 0. 0. 0. 0. T	19303	1.8%	39.0%
T. 0. 0. 0. 0. 0. 0	14135	1.3%	40.3%
0. 0. 0. 0. 0. TT. 0	12716	1.2%	41.5%
0. TT. 0. 0. 0. 0. 0	12369	1.1%	42.7%
0. TT. TT. TT. TT. TT. 0	11152	1.0%	43.7%
0. 0. TT. 0. 0. 0. 0	10197	0.9%	44.6%
0. 0. 0. 0. TT. 0. 0	9969	0.9%	45.6%
0. 0. 0. TT. 0. 0. 0	9699	0.9%	46.5%
0. TB. 0. 0. 0. 0. 0	9096	0.8%	47.3%
0. 0. 0. 0. 0. TB. 0	9051	0.8%	48.2%
0. 0. 0. 0. 0. 0. TT	8938	0.8%	49.0%
0. 0. TB. 0. 0. 0. 0	8274	0.8%	49.8%
0. 0. 0. TB. 0. 0. 0	8050	0.7%	50.5%

The CTA AFC system enables us to associate multiple trip segments with one person if they are paid by the same farecard. However, in CTA one person may use more than one farecard during a day or a week, so there could be a mismatch between the trip segment chain of a farecard and the chain of a person: a farecard's chain is either equal to or more fragmented than a person's chain. The extent of the mismatch depends on the length of the trip chain analysis period relative to the lifetime of the farecard. The longer the

analysis period, the more likely travelers have used multiple farecards, and the more fragmented the farecard's chain than the person's chain. The lifetime of the farecard could be one trip, one day, one week or even one month, but it averages 32.5 hours with the standard deviation of 46 hours[7].

[7] This is calculated based on one week AFC data from Jan. 11-17th, 2004. The mean is slightly underestimated because when a farecard exists beyond the one-week time, it is counted as one week long.

Chapter Four: Rail Path Choice Decision Modeling

This chapter presents the second case study of the utilization of ADC systems: modeling the rail path choice behavior of public transit riders. Individuals' rail path choices cumulatively determine the spatial distribution of the demand in a rail network. If the behavioral decision rules can be quantified in a systematic way, they can be used to forecast the spatial distribution of the traffic demand in the future after major network configuration changes, which are the basis to many transportation planning decisions.

This study takes advantage of the discrete choice analysis method to characterize rail riders' path choice behavior. The estimation and interpretation of the path choice models are based on the CTA rail system and the data come from two sources: the rail trip OD matrix inferred as described in Chapter Three and the attributes of alternative paths calculated from a network representation in TransCAD. This case study demonstrates that a rigorous travel behavior analysis can be performed based on data from ADC systems.

Section 4.1 describes the theoretical background of the discrete choice analysis method and introduces the topic of path choice in the CTA rail network. Section 4.2 describes the data preparation and the variable generation. Section 4.3 presents the specification and interpretation of a series of models with increasing sophistication and explanatory power,

and an application of the model to assess the impact of a hypothetical network change on the market shares of alternative paths is shown at the end of the chapter.

4.1 Rail Path Choice

Modeling travel behavior is a key aspect of transportation demand analysis wherein aggregate demand is understood to be the accumulation of individual decisions. An individual's travel decision can be generalized as a two-stage process, the long-term decisions such as residential location, work location, and car ownership; and the short-term decisions such as choice of traveling or not, choice of travel mode, choice of departure time and choice of path. An individual's short-term decisions are conditional on his long-term decisions. This case study focuses on the last stage in the decision sequence—travelers' path choice decisions in a public transit context, i.e., how public transit users make behavioral decisions over potential path choices on the rail network.

4.1.1 Discrete Choice Modeling Methods

The path choice problem involves finding the chosen path given a transportation network and a specific origin and destination. This study uses both Multinomial Logit (MNL) models and Mixed Logit models to examine the path choice problem.

The MNL model is based on random utility theory. Specifically, the utility that an individual n associates with alternative i in the choice set C_n is given by $U_{in} = V_{in} + \varepsilon_{in}$, where V_{in} is the deterministic (or systematic) part of the utility, and ε_{in} is the random term, capturing the error. The alternative with the highest utility is chosen. Therefore, the probability that alternative i is chosen by decision-maker n from choice set C_n is,

$$P(i \mid C_n) = P[U_{in} \geq U_{jn} \forall j \in C_n] = P(U_{in} = \max_{j \in C_n} U_{jn}) \quad (4\text{-}1)$$

The MNL model is derived from the assumption that the error terms of the utility function are independent and identically Gumbel distributed resulting in the following probability that an individual n chooses alternative i within the choice set C_n (Ben-Akiva and Lerman 1985) :

$$P(i \mid C_n) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \qquad (4\text{-}2)$$

Mixed Logit is a highly flexible model. It obviates the limitation of standard MNL model by allowing for random taste variation, and unrestricted substitution patterns. Mixed Logit probabilities are the integrals of standard logit probabilities over a distribution of parameters, expressed in the form,

$$P(i \mid C_n) = \int P(i \mid C_n, \beta) f(\beta) d\beta, \qquad (4\text{-}3)$$

where $P(i \mid C_n, \beta)$ is the standard MNL probability evaluated at parameter β,

$$P(i \mid C_n, \beta) = \frac{e^{V_{in}(\beta)}}{\sum_{j \in C_n} e^{V_{jn}(\beta)}}$$

and $f(\beta)$ is a density function. $V_{in}(\beta)$ is the observed portion of the utility, which depends on the parameters β. (Train 2003)

This study utilizes a binary Logit model to represent travelers' path choice behavior and uses the Mixed Logit model to test the taste variation among travelers with respect to the level of service variables such as in-vehicle travel time, the number of transfers and the transfer walk time.

4.1.2 Path Choices in CTA Rail System

CTA Rail Network and the "Loop"

This study is based on the Chicago Transit Authority rail network (see Figure 4-1), which consists of: seven rail transit lines and 143 rail stations. 1,190 rapid transit cars operate over 222 miles of track. CTA trains provide about 500,000 customer trips each day.

Figure 4-2 shows the elevated lines in the Chicago downtown area known as the Loop. Four rail lines operate on shared tracks on the Loop: the Green Line passes through the Loop linking its west and south segments; the Brown Line, Orange Line and Purple Line use the Loop to reverse direction and provide downtown passenger distribution.
On the loop, the Brown Line and Green Line North Bound travel clockwise, while the Orange Line, Purple Line and Green Line South Bound travel counter-clockwise. While the Blue Line and Red Line also pass through this area they do not use the Loop tracks. The Loop is the area where most service lines meet and where most passengers transfer between routes. The overlapping and inter-connection of rail lines in the Loop area result in multiple path choices for many OD pairs.

Figure 4-1 CTA Rail System Map, Source: Chicago Transit Authority

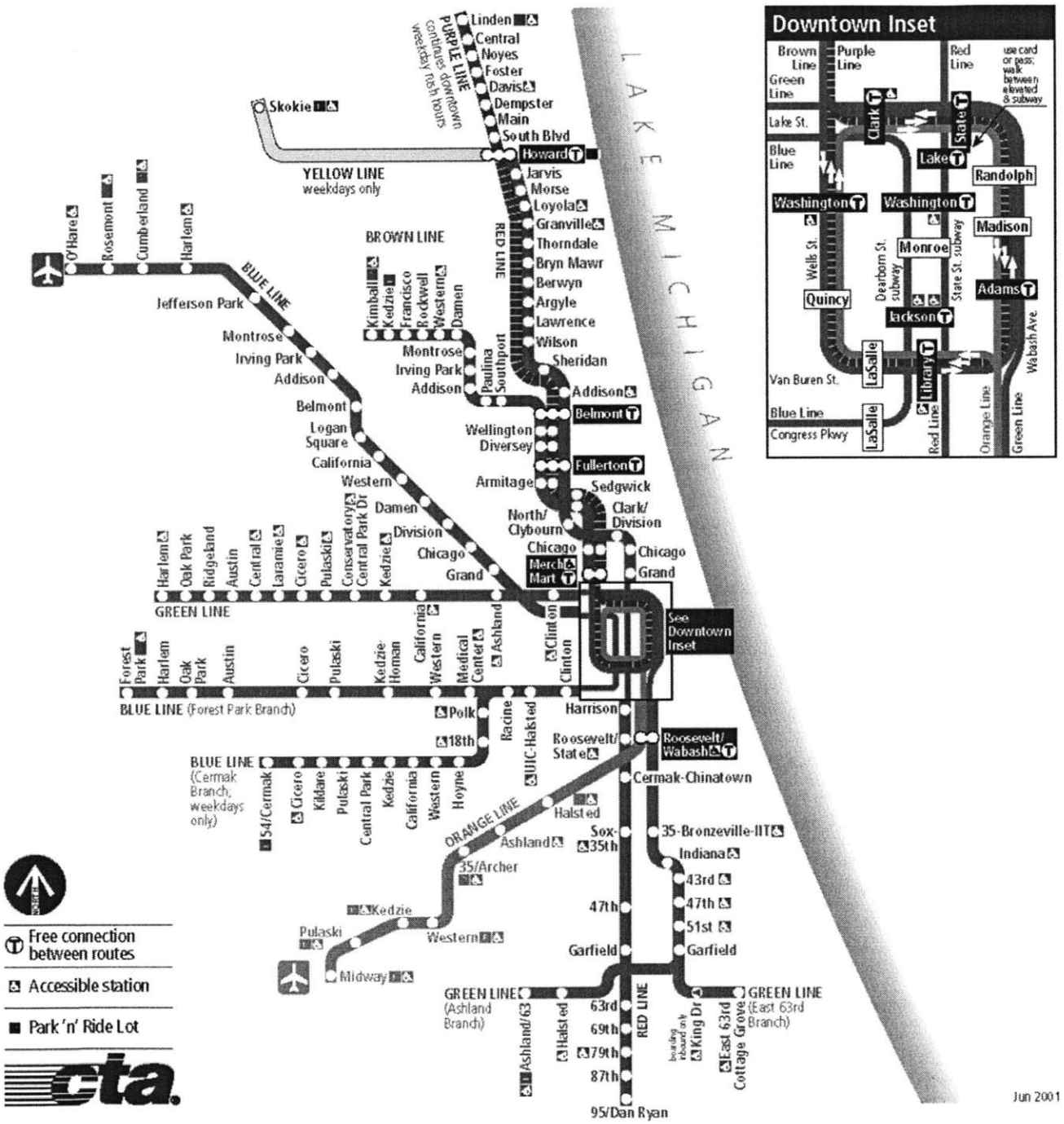
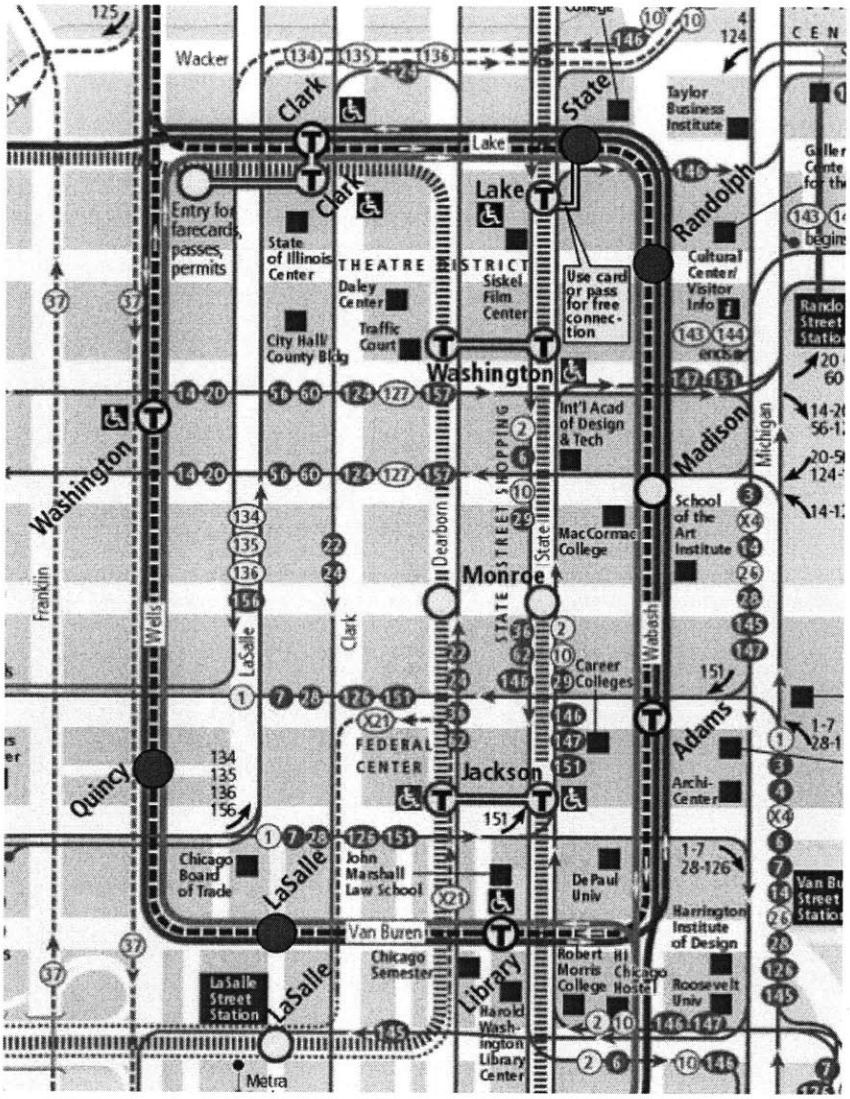


Figure 4-2 The "Loop", Source: Chicago Transit Authority



● Stations with mutually exclusive entrances for trips going in opposite directions

Path Choice Examples

Because of the overlapping and the interconnection of rail lines, travelers have choices of alternative paths traveling between certain origins and destinations. For example, travelers from Quincy to Belmont have two choices: boarding at Quincy Northbound on the Purple Line or boarding at Quincy Southbound on the Brown Line. Table 4-1 shows a

sample of the trip counts going north bound (NB) vs. south bound (SB) starting from Quincy in the PM peak between Jan. 11-17th, 2004.

Table 4-1 NB and SB Trip Counts from Quincy Station

ORIG	DEST	Total Trips	NB Trips	SB Trips	NB Trip %	SB Trip %
Quincy/Wells	Diversey	1182	1036	146	87.6%	12.4%
Quincy/Wells	Fullerton	1160	980	180	84.5%	15.5%
Quincy/Wells	Belmont-North Main	1039	834	205	80.3%	19.7%
Quincy/Wells	Armitage	857	762	95	88.9%	11.1%
Quincy/Wells	Sedgwick	715	636	79	89.0%	11.0%
Quincy/Wells	Wellington	660	541	119	82.0%	18.0%
Quincy/Wells	Southport	970	461	509	47.5%	52.5%
Quincy/Wells	Addison-North Main	422	334	88	79.1%	20.9%
Quincy/Wells	Halsted-Midway	333	326	7	97.9%	2.1%
Quincy/Wells	Howard	312	302	10	96.8%	3.2%
Quincy/Wells	Skokie	236	231	5	97.9%	2.1%
Quincy/Wells	Sheridan	258	217	41	84.1%	15.9%
Quincy/Wells	Paulina	504	190	314	37.7%	62.3%
Quincy/Wells	Chicago/Franklin	178	156	22	87.6%	12.4%
Quincy/Wells	Irving Park-Ravenswood	493	134	359	27.2%	72.8%
Quincy/Wells	Addison-Ravenswood	404	119	285	29.5%	70.5%
Quincy/Wells	Bryn Mawr	121	95	26	78.5%	21.5%
Quincy/Wells	Western-Ravenswood	530	94	436	17.7%	82.3%
Quincy/Wells	Roosevelt/Wabash	107	82	25	76.6%	23.4%
Quincy/Wells	Montrose-Ravenswood	338	78	260	23.1%	76.9%
Quincy/Wells	Damen-Ravenswood	198	57	141	28.8%	71.2%

A more careful examination of alternative travel paths reveals that the choices that passengers have vary depending on the time of day for some Origin-Destination pairs. The OD pairs are classified into two groups.

OD Pair Group One: travelers on these OD pairs have two alternatives in their path choice set and the choice set is constant throughout the day. For example, for people who go from LaSalle to all Orange line stations outside the Loop, the choice set is constant as follows,

1: board westbound on the Orange line

2: board eastbound on the Brown line; transfer at Adams to the Orange line

OD Pair Group Two: for travelers on these OD pairs, the choice set changes by time period depending on whether a particular line is running or not. For example, for trips from Quincy to Fullerton, the choice set depends on whether the Purple Line is running. When the Purple Line is running, there are three alternatives,

1: board northbound on the Purple line

2: board northbound on the Orange line; transfer at Clark to the Brown line

3: board southbound on the Brown line

When the Purple line is not running, only alternatives two and three are available. The variation of the path choice set complicates path choice modeling and to avoid this, this study only considers the PM peak period, when all services lines are running and the path choice sets are fixed.

Stations with Exclusive Entrances for Each Direction

To estimate path choice models, it is necessary to determine which path an individual actually takes. In general the CTA path choice from origins in the Loop is not automatically observable since passengers can enter through any entrance turnstile and travel in either direction. Fortunately, four stations in the Loop (Quincy, LaSalle, Randolph, and State/Lake) have mutually exclusive entrances for each direction. Thus a rider's choice of entrance (recorded by the AFC system) at these four stations indicates the direction the rider chooses. The simplicity of the CTA rail network allows us to infer the exact path travelers choose from the alternatives by using this direction information. The four stations of interest are shown in Figure 4-2.

4.2 Data Preparation

Two data sets are used in this study: the rail origin-destination matrix described in chapter three, and the attributes of alternative paths. The CTA rail system is modeled in TransCAD and the attributes of each alternative path are determined from the network representation.

4.2.1 Rail OD Matrix

In Chapter Three, a one-week rail trip OD matrix was estimated by applying the rail destination inference method. The path choice model utilizes a subset of the rail OD matrix defined as follows:

1) origins at any of the three stations: Quincy, LaSalle, and Randolph[8]
2) destinations outside the loop
3) ODs that have practical alternative paths.
4) between 4 and 7pm on weekdays

There are 29137 recorded rail trips originating at Quincy, LaSalle, and Randolph to destinations outside the Loop in the PM peak hours from Jan.12-16th, 2004. They comprise 502 OD pairs, of which 402 have practical path alternatives which can be used in this path choice modeling application. Because all trips on the same OD pair have the same attributes for each alternative, they are aggregated into 402 distinct observations with the number of trips between the OD pair as the weight of each observation.

8 Trips from State/Lake are not used because these trips have a more complex choice set.

4.2.2 Rail Network Representation and Path Attributes Calculation

The CTA rail network is modeled in TransCAD using the "Route System" and "Transit Network" tools and attributes of alternative paths are calculated using the "Transit Skimming" function.

The three major data types required for the TransCAD network model are, train run time on each track segment, the service frequency on each route, and the transfer walk time between connecting routes.

The transfer walk time came from direct observation by the author in Chicago. The transfers are categorized into four groups: cross-platform transfer, overpass or underpass transfer, walkway (Jackson street and Washington street between the Red Line and Blue Line), and Loop-subway transfer, and the transfer walk time for each are shown in Table 4-2. There are three cases for the Loop-subway transfer depending on the transfer location.

Table 4-2 Rail to Rail Transfer Categories

Transfer Type	**Direction**	**Transfer walk time (minutes)**
Same Platform	both	0.1
Overpass/Underpass	both	0.9
Walkway (Jackson, Washington)	both	2.7
Loop-Subway		
At Clark/Lake	Loop to Subway	3.5
	Subway to Loop	3.1
At State/Lake	Loop to Subway	3.3
	Subway to Loop	3.3
At Roosevelt	Loop to Subway	3
	Subway to Loop	3

The train service frequency is derived from the train schedules between 4 and 7pm along each line and branch in each direction. For some routes such as the Green Line (see Figure 4-3), the headways are quite stable during the PM rush hours. For other routes such as the Red Line NB (see Figure 4-4), the headways fluctuate. In both cases however, the mean headway between 4 and 7pm is used to characterize frequencies. The running time on each track segment is also derived from the train schedules.

Figure 4-3 Green Line SB Weekday Headway 4 to 7pm

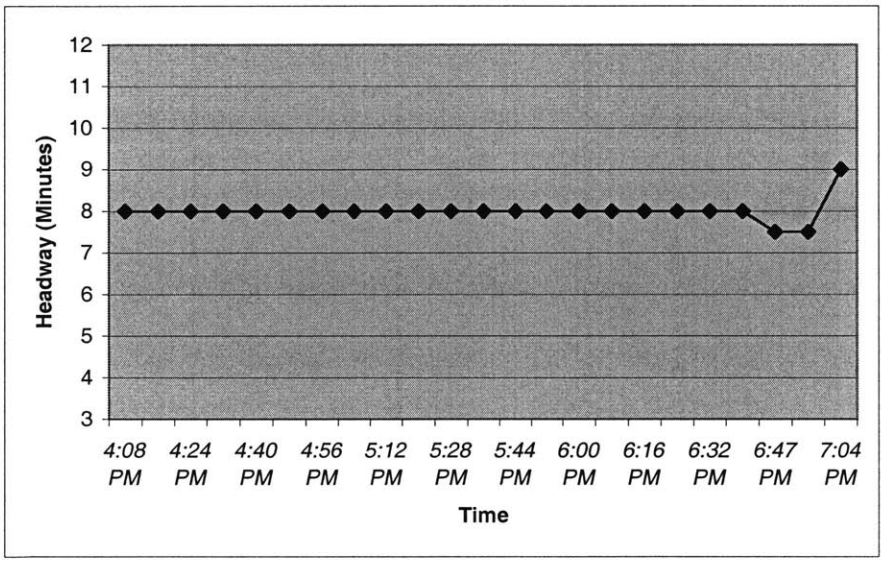
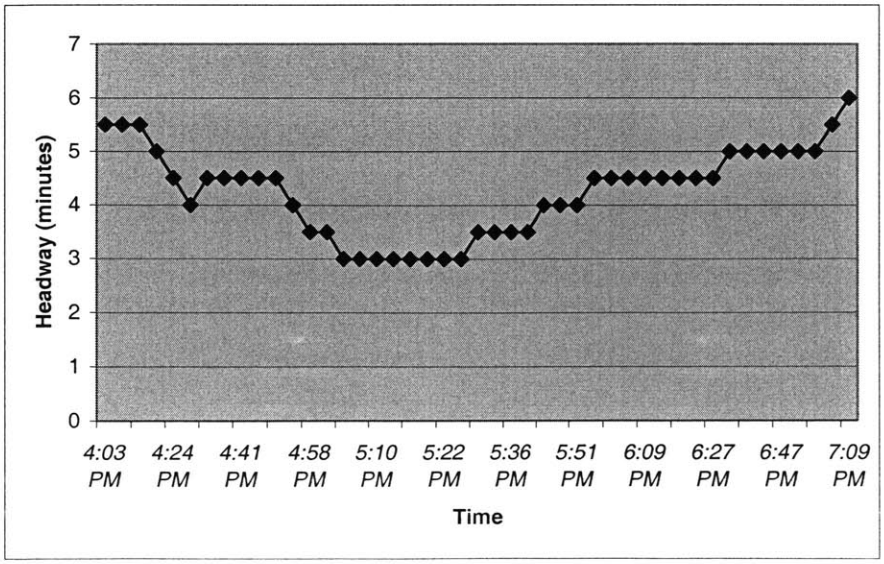


Figure 4-4 Red Line NB Weekday Headway 4 to 7pm



Within the TransCAD model of the CTA rail network, attributes of the alternative paths are calculated using the "Transit Skim" function. The OD pairs of interest in this study are those between Quincy, LaSalle, and Randolph and stations outside the loop. The path direction from the origin station is enforced so that for OD pairs with alternative paths, the attributes of the paths going clockwise and counter-clockwise can be distinguished.

The output of the TransCAD model is a series of matrices, each of which corresponds to one trip attribute for the OD pairs of interest.

4.2.3 Variable Generation

The rail trip OD data are joined with path attribute data to generate the eight independent variables[9] listed in Table 4-3.

Table 4-3 Independent Variables

Variable Name	Description	Units
IVT	In-Vehicle Travel Time	Minutes
WAITA	Initial Waiting Time	Minutes
WAITB	Transfer Wait Time*	Minutes
WALKING	Transfer Walk Time*	Minutes
TFR	Number of Transfers	Units
LENGTH	Total Track Length	Miles
STOPS	Number of Stops	Units
TOTAL	Total Travel Time	Minutes

* The wait time is defined as half the mean headway.

These variables are categorized into three groups, the base variables, the transfer related attributes, and other trip attributes as shown in Figure 4-5.

[9] Two additional variables are available but not included as independent variables. First the number of trips on each OD pair enters the model as a weight. Second, the total dwell time is not included because the dwell time at each station is assumed to be constant so that the total dwell time is perfectly correlated with the number of stops.

Figure 4-5 Independent Variable Categories

Base Variables	Transfer Attributes	Other Trip Attributes
In-Vehicle Travel Time	Transfer Waiting Time	Total Travel Time
Initial Waiting Time	Transfer Walking Time	Total Track Length
Number of Transfers		Number of Stops

Discrete choice model estimation requires not only the attributes of the selected choice but also the attributes of the non-selected choices and so the attributes of the non-selected path are also calculated. For this binary choice model, there are two sets of attributes for each observation.

For the OD pairs with three alternative paths, the two alternatives going in the same direction are combined to form one composite alternative—the attributes of the combined path are used in the model. Specifically, the initial wait time of the combined path is calculated as:

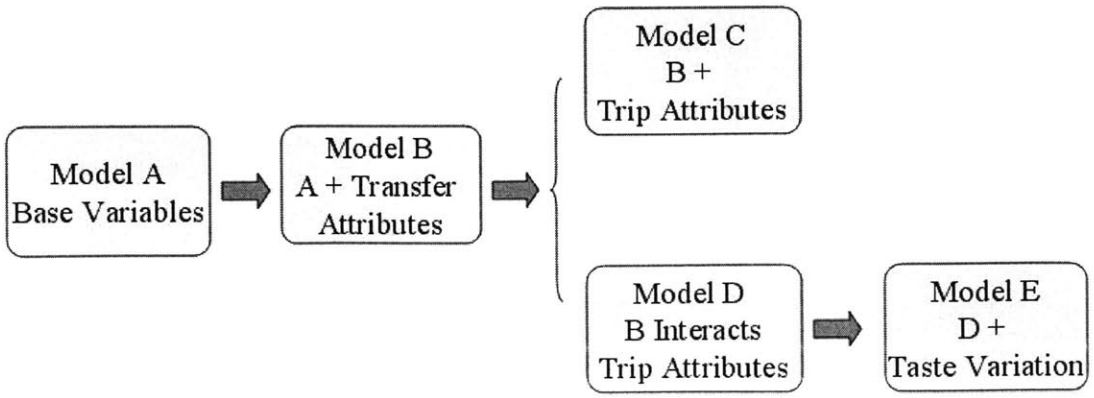
$$w = \frac{1}{\frac{1}{w_1} + \frac{1}{w_2}}$$

where w is the initial wait time for the composite alternative, and w_1 and w_2 are the initial wait times of the two original alternatives.

4.3 Model Estimation and Interpretation

This section presents a series of models of the rail path choice decision. The models are developed with increasing sophistication and explanatory power. As shown in Figure 4-6, the series starts with the simplest model (Model A) which includes only the base variables, in-vehicle travel time (IVT), number of transfers (TFR) and initial wait time (WAITA). The simplest model checks the significance of the three explanatory variables and evaluates the tradeoffs among them. Model B introduces transfer wait time and transfer walk time into the model to characterize the transfer attributes. Part of the transfer effect can be captured by the two transfer attribute variables. Both Models C and D include other trip attributes such as total trip length, total travel time, and number of stops in the model, but in different ways. Model C introduces the new variables in parallel with the existing variables in Model B, whereas Model D uses interactions between the new variables and the existing variables. As will be shown Model D works much better than Model C so Model D is chosen to be the base model for Model E, in which random coefficients are introduced to test travelers' taste variation over different level of service (LOS) attributes.

Figure 4-6 The Model Specification Development



4.3.1 Simplest Model (Model A)

In the simplest model specification, the utility function only includes in-vehicle travel time (IVT), initial wait time (WAITA), and number of transfers (TFR). This model checks the significance of the three base variables and the tradeoffs among them.

Table 4-4 Model A Estimation Result

Model A

Independent Variable	Estimated Coefficient	Standard Error	T-Test
In-Vehicle Travel Time	-0.280	0.068	**-4.089**
Number of Transfers	-2.526	0.306	**-8.266**
Initial Waiting Time	0.032	0.090	0.349

Auxillary Statistics

Number of estimated parameters	3
Null log-likelihood:	-172.926
Final log-likelihood:	-109.252
Likelihood ratio test:	127.348
Rho-square:	0.368
Adjusted Rho-Square	**0.351**

The estimation results of Model A (Table 4-4) show that IVT and TFR are highly significant and both have the expected signs. The negative signs mean that, all else being equal, travelers tend to choose the path with shorter in-vehicle travel time and fewer transfers.

The estimated utility function is $U_i = -0.280 * IVT_i - 2.526 * TFR_i + \varepsilon_i$. Only the relative magnitudes between variable coefficients have an interpretable meaning. Specifically the ratio between the coefficients of IVT and TFR is the marginal rate of substitution between the two variables that measures the tradeoff between the number of transfers and the in-vehicle travel time. In Model A, one transfer is perceived by a typical traveler as

being equivalent to 9.04 minutes of in-vehicle travel time, calculated by the following equation.

$$MRS = \frac{\partial U / \partial TFR}{\partial U / \partial IVT} = \frac{-2.526}{-0.280} = 9.04$$

The t-test for initial wait time (WAITA) is low indicating that in the situation being studied the initial wait time is not important in the typical traveler's path choice decision. This may be because there is little variation in the initial wait time difference between alternative paths in the CTA rail network in the PM peak period: the standard deviation of the waiting time difference is 2.4 minutes.

ρ^2 is an informal goodness of fit index for discrete choice models that measures the fraction of the initial log-likelihood value explained by the model: it is analogous to R^2 in linear regression models. The adjusted ρ^2 is similar to ρ^2, but corrected for the number of parameters estimated (Ben-Akiva 1985). Usually a value between 0.3 and 0.4 for the adjusted ρ^2 is a good result for a model specification, which is equivalent to an adjusted R^2 of 0.7 to 0.85 for linear regression models (Chu 2002). Model A yields an adjusted ρ^2 of 0.351, indicating reasonably good explanatory power for the model. When more sophisticated models are formulated, adjusted ρ^2 is expected to increase.

The results of Model A are compared to the previous transfer penalty study by Guo (2003) based on the Massachusetts Bay Transportation Authority (MBTA) rail system (see Table 4-5).

Table 4-5 Comparison between Previous Transfer Penalty Study and Model A[10]

Model	Model B-1 in Guo 2003	Model A
Variables in Utility Function	Transfers	Transfers
	In-vehicle Travel Time	In-vehicle Travel Time
	Access Walking Time	*Zero Difference*
	Zero Difference	Initial Waiting Time
Systems and Models	MBTA Subway Path Choice	CTA Rail Path Choice
Transfer Penalty Measured in In-Vehicle Travel Time	10.6 minutes/transfer	9.04 minutes/transfer

Both models have number of transfers (TFR) and in-vehicle travel time (IVT) as independent variables. In model A, "Access Walking Time" is always zero and in Guo's model, "initial wait time" of transfer path and non-transfer path are identical, also resulting in a zero difference. Given the differences between the MBTA and CTA rail systems the results of 10.6 minutes/transfer and 9.04 minutes/transfer are strikingly close.

4.3.2 Models with Transfer Attributes (Model B)

In Model B, the transfer wait time and the transfer walk time are included to better represent transfer attributes. The estimation results for Model B1 are shown in Table 4-6.

[10] In Guo's study, the transfer penalty is measured in terms of access walking time. The author recalculated the transfer penalty as measured by in-vehicle travel time (10.6 minutes/transfer) based on Guo's model estimation results.

Table 4-6 Model B1 Estimation Result

Model B1

Independent Variable	Estimated Coefficient	Standard Error	T-Test
In-Vehicle Travel Time	-0.313	0.091	**-3.444**
Number of Transfers	-2.157	1.236	**-1.746**
Initial Waiting Time	-0.048	0.145	-0.330
Transfer Waiting Time	0.018	0.412	0.043
Transfer Walking Time	-1.065	0.582	**-1.832**

Auxillary Statistics

Number of estimated parameters	5
Null log-likelihood:	-172.926
Final log-likelihood:	-107.447
Likelihood ratio test:	130.959
Rho-square:	0.379
Adjusted Rho-Square	**0.350**

While the ρ^2 increases from 0.368 to 0.379 in Model B1, the adjusted ρ^2 is virtually identical, indicating no better data fit than in Model A.

Neither the initial wait time nor the transfer wait time are significant for the reason discussed in Model A.

The transfer walk time is significant and its negative sign suggests travelers prefer paths with shorter transfer walk time, all else being equal. The absolute value of its coefficient (1.065) is much greater than that of the in-vehicle travel time (0.313). This means that travelers dislike transfer walk time much more than in-vehicle travel time. The tradeoff between the transfer walk time and the in-vehicle travel time is measured by their marginal rate of substitution, 3.4 as calculated below, which means that 3.4 minutes of in-vehicle travel time is perceived as the same as one minute of transfer walk time.

$$MRS = \frac{\partial U / \partial WALK}{\partial U / \partial IVT} = \frac{-1.065}{-0.313} = 3.4$$

The ratio between the coefficients of in-vehicle travel time and number of transfers decreases to 6.88, compared with 9.04 in Model A. This is because part of the transfer penalty is now explained by transfer attribute variables, specifically the transfer walk time variable.

This result is compared to Model B-2 in Guo (2003) in Table 4-7.

Table 4-7 Comparison between Model B2 in Guo 2003 and Model B1 in this Research

Model	**Model B-2 in Guo 2003**	**Model B1**
Variables in Utility Function	Transfers	Transfers
	In-vehicle Travel Time	In-vehicle Travel Time
	Transfer walk time	Transfer walk time
	Transfer wait time	Transfer wait time
	Access Walk Time	*Zero Difference*
	Zero Difference	Initial Wait Time
Systems	MBTA	CTA
$\frac{\partial U / \partial TFR}{\partial U / \partial IVT}$	5.27 minutes/transfer	6.88 minutes/transfer
$\frac{\partial U / \partial WALK}{\partial U / \partial IVT}$	5.8 minutes/minute	3.4 minutes/minute

Both models have similar independent variables: number of transfers (TFR), in-vehicle travel time (IVT), transfer walk time (WALK), and transfer wait time (WAITB). The ratios between TFR and IVT from both models are similar (5.27 versus 6.88 minutes/transfer), while the ratios between WALK and IVT for both models are not as close. The difference between 5.8 in Boston versus 3.4 in Chicago may be because the transfer facilities (corridor, overpass or underpass) are better designed in Chicago than in

Boston so that travelers' dislike of transfer walk time relative to in-vehicle travel time in Chicago is less than in Boston.

Since the transfer wait time and the initial wait time are far from significant, these two variables are dropped from Model B1 to get a variation, Model B2. The estimation result of Model B2 is given in Table 4-8. The adjusted ρ^2 goes up from 0.350 to 0.361, which indicates that dropping the transfer wait time and the initial wait time simplifies the model without significantly degrading the explanatory power of the model. Subsequent model developments are based on Model B2.

Table 4-8 Model B2 Estimation Result

Model B2

Independent Variable	Estimated Coefficient	Standard Error	T-Test
In-Vehicle Travel Time	-0.290	0.057	**-5.078**
Number of Transfers	-2.108	0.377	**-5.591**
Transfer Walking Time	-0.936	0.510	**-1.835**

Auxillary Statistics

Number of estimated parameters	3
Null log-likelihood:	-172.926
Final log-likelihood:	-107.567
Likelihood ratio test:	130.719
Rho-square:	0.378
Adjusted Rho-Square	**0.361**

4.3.3 Models with other Trip Attributes (Models C and D)

Travelers' perceptions of IVT, TFR, WAIT, WALK etc. may be affected by other trip attributes such as total trip length, total travel time, and number of stops. Based on Model B2, these variables are included in the specifications of Models C and D in two different ways. Model C introduces the trip attributes in parallel with the base variables while

Model D includes interaction terms combining these variables with some of the base variables.

Table 4-9 reports the estimation results of Model C. The adjusted ρ^2 drops from 0.361 to 0.341, which indicates that the parallel addition is not effective. The track length and the number of stops do not show up as significant in Model C. This is reasonable because both the track length and the number of stops are highly positively correlated with the in-vehicle travel time. Adding these two variables does not improve the explanatory power of the model.

Table 4-9 Model C Estimation Result

Model C			
Independent Variable	**Estimated Coefficient**	**Standard Error**	**T-Test**
In-Vehicle Travel Time	-0.688	0.460	**-1.494**
Track Length	0.325	0.767	0.423
Number of Stops	0.272	0.289	0.940
Number of Transfers	-2.150	1.240	**-1.734**
Initial Waiting Time	0.124	0.211	0.587
Transfer Waiting Time	0.007	0.412	0.017
Transfer Walking Time	-1.148	0.592	**-1.937**
Auxillary Statistics			
Number of estimated parameters		7	
Null log-likelihood:		-172.926	
Final log-likelihood:		-107.002	
Likelihood ratio test:		131.848	
Rho-square:		0.381	
Adjusted Rho-Square		0.341	

In Model D the new variables are introduced as four interaction terms generated by multiplying IVT and TFR with trip length (LENGTH) and number of stops (STOPS). The specification and estimation results of Model D1 are given in Table 4-10.

The effects of the four interaction terms are not identical. The two IVT interaction terms (IVT*LENGTH and IVT*STOPS) are significant at the 10% level, suggesting that the marginal utility of in-vehicle travel time varies with trip length and number of stops. However, the interaction terms for TFR (TFR*LENGTH, TFR*STOPS) are not significant, which indicates that the marginal utility of TFR is not related to trip length or number of stops. Consequently, the two interaction terms for TFR are dropped from Model D1 to obtain Model D2. As shown in Table 4-11, Model D2 is improved from Model B2 with an increase of adjusted ρ^2 from 0.361 to 0.371.

Table 4-10 Model D1 Estimation Results

Model D1

Independent Variable	Estimated Coefficient	Standard Error	T-Test
IVT	-0.757	0.185	**-4.094**
IVT*LENGTH	0.019	0.011	**1.714**
IVT*STOPS	0.007	0.004	**1.808**
TFR	-1.578	1.092	-1.445
TFR*LEN	0.065	0.098	0.662
TFR*STOPS	-0.032	0.075	-0.428
WALK	-1.727	0.713	**-2.421**

Auxillary Statistics

Number of estimated parameters	7.000
Null log-likelihood:	-172.926
Final log-likelihood:	-103.524
Likelihood ratio test:	138.804
Rho-square:	0.401
Adjusted Rho-Square	**0.361**

Table 4-11 Model D2 Estimation Results

Model D2

Independent Variable	Estimated Coefficient	Standard Error	T-Test
IVT	-0.735	0.178	**-4.122**
IVT*LENGTH	0.017	0.011	1.613
IVT*STOPS	0.007	0.004	**1.789**
TFR	-1.569	0.424	**-3.700**
WALK	-1.497	0.574	**-2.610**

Auxillary Statistics

Number of estimated parameters	5.000
Null log-likelihood:	-172.926
Final log-likelihood:	-103.744
Likelihood ratio test:	138.365
Rho-square:	0.400
Adjusted Rho-Square	**0.371**

A model with interaction term TFR*TFR was also tested to see if travelers have a non-linear perception of transfers, but this did not turn out to be significant.

Because of the interaction terms, interpretation of the coefficients is more complicated. First, the marginal utility of IVT is no longer a constant but a function of LENGTH and STOPS:

$$\partial U / \partial IVT = -0.735 + 0.017 * LENGTH + 0.007 * STOPS .$$

Therefore depending on the trip length and the number of stops, the marginal utility of IVT will vary. Table 4-12 shows the number of stops and trip length for the shortest trip, the average trip and the longest trip, and the corresponding marginal utilities of IVT. The average trip length is 7.8 miles with 13 stops and a corresponding marginal utility is -0.51. The longer the trip, the less travelers care about one additional minute of IVT.

Table 4-12 Stops and Trip Length for the Shortest, the Average and the Longest Trips

	Shortest Trip	Average Trip	Longest Trip
Length (Miles)	0.70	7.80	18.00
Stops	2.00	13.00	26.00
dU/dIVT	-0.71	-0.51	-0.25

Because of the dependence of marginal utility of in-vehicle travel time on trip length and number of stops, the marginal rate of substitution (MRS) between number of transfers and in-vehicle travel time is also a function of these variables:

$$MRS = \frac{\partial U / \partial TFR}{\partial U / \partial IVT} = \frac{-1.569}{-0.735 + 0.017 * LENGTH + 0.007 * STOPS}$$

Similarly the marginal rate of substitution between transfer walk time and in-vehicle travel time is a function of trip length and the number of stops

$$MRS = \frac{\partial U / \partial WALK}{\partial U / \partial IVT} = \frac{-1.497}{-0.735 + 0.017 * LENGTH + 0.007 * STOPS}$$

The values for the two marginal rates of substitution are calculated for the three cases, the shortest, the average and the longest trips, as shown in Table 4-13. For the average case, the tradeoff between TFR and IVT is 3.07 minutes/transfer and one minute of transfer walk time is equivalent to 2.93 minutes of in-vehicle travel time.

Table 4-13 MRS between TFR and IVT, WALK and IVT for Different Trip Lengths

	Shortest Trip	Average Trip	Longest Trip
(dU/dTFR) /(dU/dIVT)	2.21	3.07	6.36
(dU/dWALK) /(dU/dIVT)	2.11	2.93	6.07

Table 4-14 summarizes adjusted ρ^2, marginal rates of substitution between TFR and IVT, and between WALK and IVT for Models A through D.

Table 4-14 A Summary of Models A through D

Model	Variables in Utility Function	Adjusted ρ^2	$\frac{\partial U / \partial TFR}{\partial U / \partial IVT}$	$\frac{\partial U / \partial WALK}{\partial U / \partial IVT}$
A	In-Vehicle Travel Time Number of Transfer Initial Wait Time	0.351	9.04minutes/tfr	N/A
B2	In-Vehicle Travel Time Number of Transfer Transfer walk time	0.361	7.26 minutes/tfr	3.23
D2	In-Vehicle Travel Time Number of Transfer Transfer walk time IVT*LENGTH IVT*STOPS	0.371	3.07* minutes/tfr	2.93*

* These values are calculated for the case with average trip length, and average number of stops.

The values of $\frac{\partial U / \partial WALK}{\partial U / \partial IVT}$ are relatively stable across models at about 3 minutes of IVT per minute of transfer walk time. The values of $\frac{\partial U / \partial TFR}{\partial U / \partial IVT}$ range from 3.07 minutes/transfer to 9.04 minutes/transfer, depending on the model specification. Therefore caution is needed when applying these values. Based on the model application context, specifically the data availability for the independent variables, the appropriate model should be selected and the corresponding coefficients used.

Model D2 is the best MNL model specification and is used as the base for the Mixed Logit model estimation in the following section.

4.3.4 Models with Taste Variation (Model E)

Model E introduces random coefficients to test the travelers' taste variation. For three variables—in-vehicle travel time, number of transfers and transfer walk time, the coefficients are relaxed from a constant to a distribution. The model estimates both the mean values and the standard deviations of the coefficients given the assumed normal distribution.

Mixed logit models are generally estimated by simulation (Train 2003). In this study, simulations were performed in BIOGEME 0.7 using 100, 500, 1000, 2000, 3000, 4000, and 5000 Halton draws. The simulation estimation results are given in Table 4-15. The ρ^2, the initial log-likelihood, final log-likelihood, log-likelihood ratio test, and the variables coefficients are stable after 4000 Halton draws. The estimation results of the model with 5000 Halton draws are presented in Table 4-16.

Table 4-15 Simulation Estimation Results of Model E with 100, 500, 1000, 2000, 3000, 4000, and 5000 Halton Draws

Halson Draw	100	500	1000	2000	3000	4000	5000
Null Log-Likelihood	-172.926	-172.926	-172.926	-172.926	-172.926	-172.926	-172.926
Final Log-Likelihood	-103.121	-103.093	-103.076	-103.077	-103.080	-103.079	-103.074
Log Likelihood Ratio Test	139.610	139.666	139.701	139.698	139.692	139.695	139.705
Rho-Square	0.4037	0.4038	0.4039	0.4039	0.4039	0.4039	0.4039
IVT	-0.847	-0.853	-0.861	-0.859	-0.859	-0.862	-0.863
IVTLEN	0.017	0.017	0.017	0.017	0.017	0.017	0.017
IVTSTOPS	0.007	0.007	0.008	0.008	0.008	0.008	0.008
IVT_S	0.239	0.253	0.271	0.267	0.267	0.272	0.274
TFR	-1.721	-1.751	-1.761	-1.756	-1.757	-1.758	-1.760
TFR_S	0.126	0.025	0.031	0.039	0.038	0.048	0.044
WALK	-2.359	-2.392	-2.407	-2.406	-2.404	-2.410	-2.412
WALK_S	1.832	1.875	1.883	1.885	1.883	1.888	1.889

Table 4-16 Model E Estimation Results

Model E

Independent Variable	Parameter	Estimated Value	Robust Standard Error	Robust T-test
In-Vehicle Travel Time	Mean Coefficient	-0.863	0.225	**-3.826**
	Std. dev. of Coefficient	0.274	0.344	0.795
Number of Transfers	Mean Coefficient	-1.760	0.733	**-2.400**
	Std. dev. of Coefficient	0.044	0.078	0.572
Transfer Walking Time	Mean Coefficient	-2.412	1.032	**-2.338**
	Std. dev. of Coefficient	1.889	0.978	**1.932**
IVT*LENGTH	Mean Coefficient	0.017	0.010	**1.679**
IVT*STOPS	Mean Coefficient	0.008	0.003	**2.299**

Auxillary Statistics

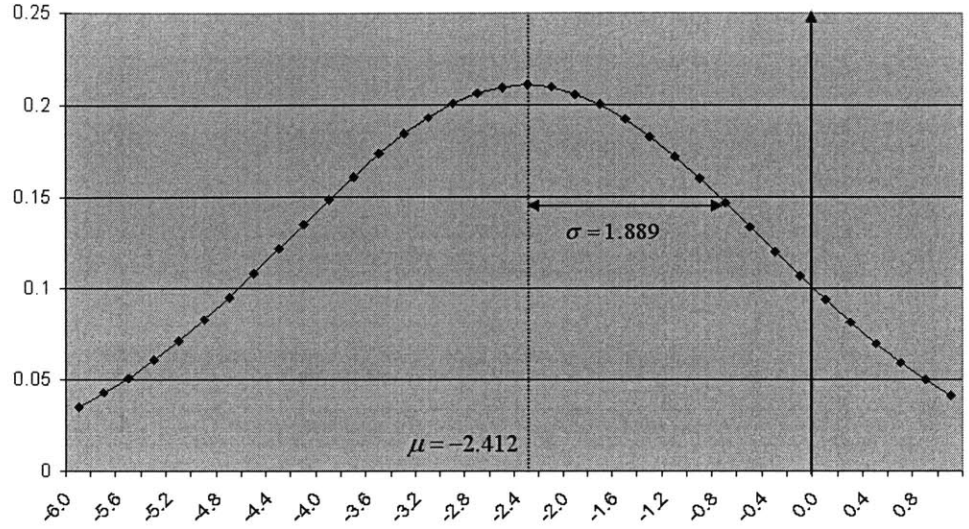
Number of Halson Draw	5000
Number of estimated parameters	8
Null log-likelihood:	-172.926
Final log-likelihood:	-103.074
Likelihood ratio test:	139.705
Rho-square:	0.404
Adjusted Rho-Square	**0.358**

The mean values of the coefficients of in-vehicle travel time, number of transfers and transfer walk time are all significant and have the expected negative signs. The interaction terms between IVT and LENGTH, between IVT and STOPS are also significant or marginally significant.

The major difference between the Mixed Logit model and the MNL model is in the standard deviation terms of the coefficients. The standard deviation of the coefficient of transfer walk time is significant at the 10% level, which indicates that this coefficient does vary in the population. Travelers have significant differences in their perceptions of transfer walk time. The difference in travelers' perception is represented by a normal

distribution with a mean of -2.412 and standard deviation of 1.889 as shown in Figure 4-7. Most travelers dislike longer transfer walk time but they dislike it to different extents.

Figure 4-7 Normal Distribution of the Coefficient of Transfer walk time.



This is to be expected since the same walk time can mean very different things for different groups of people. Seniors and passengers with large packages will strongly dislike the transfer walk, especially if there are level changes associated with it such as going between the Loop and underground stations. This is particularly true in this model specification where there are no socio-economic variables to describe the travelers' attributes such as age and trip purpose.

The standard deviation of the coefficient of the number of transfers is not significant which indicates that travelers' perceptions of the number of transfers are similar, concentrating around the mean value (-1.76), which is interpreted in the same way as in the MNL model: one additional transfer decreases travelers' utility by 1.76 units.

The non-significance of the standard deviation of the coefficient of in-vehicle travel time needs to be distinguished from the non-significance of the standard deviation of the coefficient of number of transfers. This is because of the two interaction terms IVT*LENGTH and IVT*STOPS. The non-significance of the standard deviation of in-vehicle travel time means that at the same trip length and number of stops, travelers' preferences for in-vehicle travel time are similar. In other words, the travelers' taste variations over in-vehicle travel time are mostly explained by the trip length and number of stops. We explicitly model this taste variation through the interaction terms IVT*LENGTH and IVT*STOPS.

The utility function from the Model E estimation result is:

$$U_i = -0.863 * IVT_i - 1.760 * TFR_i - 2.412 * WALK + 0.017 * IVT * LENGTH + 0.008 * IVT * STOPS + \varepsilon_i$$

The ρ^2 slightly increases from 0.400 in Model D2 to 0.404 while the adjusted ρ^2 drops, indicating that the general explanatory power of the model does not improve significantly. But Model E offers clear evidence that travelers have taste variation with respect to transfer walk time, but no taste variation with respect to the number of transfers, or in-vehicle travel time given that trip length and number of stops are controlled for.

The marginal utility of in-vehicle travel time, the marginal rates of substitution between number of transfers and in-vehicle travel time, and between transfer walk time and in-vehicle travel time are recalculated for the three trip length cases as shown in Table 4-17.

Table 4-17 Marginal Utility and MRS for the Different Trip Lengths

	Shortest Trip	Average Trip	Longest Trip
Length(Mile)	0.70	7.80	18.00
Stops	2	13	26
dU/dIVT	-0.84	-0.63	-0.35
(dU/dTFR) /(dU/dIVT)	2.11	2.80	4.97
(dU/dWALK)/ (dU/dIVT)	2.89	3.83	6.80

For the average case (a trip length of 7.8 miles with 13 stops) the marginal utility of in-vehicle travel time is -0.63. The tradeoff between TFR and IVT is 2.80 minutes/transfer and one minute of transfer walk time is equivalent to 3.83 minutes[11] of in-vehicle travel time.

Table 4-18 Comparison between the Mixed Logit Model and the MNL Model

Model		D2	E
Rho-square:		0.400	0.404
Adjusted Rho-Square		0.371	0.358
dU/dIVT	Shortest Trip	-0.71	-0.84
	Average Trip	-0.51	-0.63
	Longest Trip	-0.25	-0.35
(dU/dTFR) /(dU/dIVT)	Shortest Trip	2.21	2.11
	Average Trip	**3.07**	**2.80**
	Longest Trip	6.36	4.97
(dU/dWALK) /(dU/dIVT)	Shortest Trip	2.11	2.89
	Average Trip	**2.93**	**3.83**
	Longest Trip	6.07	6.80

Table 4-18 compares the results of the mixed logit model (Model E) with the MNL model (Model D2). The ratio between the coefficients of the number of transfers and the in-vehicle travel time is slightly smaller in the Mixed Logit model (2.80 for the average trip) than in the MNL model (3.07 for the average trip) while the ratio between transfer

[11]In Model E the transfer walk time is estimated as a normal distribution. The quotient of a normal distribution over a constant is also a normal distribution. Therefore the marginal rate of substitution between transfer walk time and in-vehicle travel time is a normal distribution. The value 3.83 in Table 4-18 is the mean value of the distribution.

walk time and in-vehicle travel time is greater in the Mixed Logit model (3.83 for the average trip) than in the MNL model (2.93 for the average case).

4.3.5 Model Application

The model estimated above can be used to forecast the probability that an individual chooses a certain rail path for a particular journey. Individuals' choices collectively determine the market shares of the alternative paths for any journey. The spatial distribution of the total demand on the rail network can be determined by aggregating the market shares of alternative paths between all OD pairs. Then the demand for each service route can be estimated accordingly. This section demonstrates how the model can be used to assess the impact of a hypothetical network change on the market shares of the alternative paths for a given journey type.

The utility function estimation above measures the tradeoffs among the factors that travelers consider in choosing their rail path. Given a choice set C_n with attributes of each of the alternatives being available, the probability that an individual n chooses alternative i is given by equation 4-2.

However, prediction for a specific individual's choice is usually of little use in a transit agency's planning or operational decision-making. Instead, decisions are made based on the forecast of aggregate demand in particular time periods, such as the number of total passenger trips along a service line in the weekday PM rush hour. Therefore a linkage

between the disaggregate level models estimated above and the aggregate forecasts of interest to transit agencies is needed if the models are to be of value.

Trips from Quincy Station to Paulina Station are taken as an example to illustrate the model application process. During the weekday PM rush hours from Jan.12-16th, 2004, there were 492 rail trips recorded from Quincy to Paulina. There are two alternative paths for this OD pair.

1: board the Brown line southbound;

2: board the Purple line northbound; transfer at Belmont to the Brown Line.

The trip attributes of the two alternatives are given in Table 4-19.

Table 4-19 Attributes of Alternative Paths from Quincy to Paulina

Orig	DEST	Path	DIR	Weight	IVT	WALK	TFR	LENGTH	STOPS
Quincy	Paulina	1	SB	311	22.96	0	0	7.11	17
Quincy	Paulina	2	NB	181	18.82	0.10	1	6.07	11

The AFC system indicates a total of 181trips took the northbound path and 311 trips took the southbound path. How will this split change if some LOS attributes change because of a new service or a new operational plan? For example, the in-vehicle travel time of alternative 2 is currently 18.82. Suppose, a new type of train will decrease the in-vehicle travel time by 10%, what is the impact on the share of trips on both paths? In another example, suppose that Belmont Station is under reconstruction requiring a detour of the transfer path between Purple Line to Brown Line. As a result the walk time for this transfer increases from 0.1 to 0.9 minute. What is the impact of this change on the market shares for the Purple and Brown lines?

The Model E specification is used in this example. The estimated utility function[12] is

$$U_i = -0.863 * IVT_i - 1.760 * TFR_i - 2.412 * WALK + 0.017 * IVT * LENGTH + 0.008 * IVT * STOPS + \varepsilon_i$$

The utility of each path can be calculated by plugging the path attribute values into the utility function. The probability that a person n will take path i is a function of both the utility of the chosen alternative and the utility of the non-chosen alternative (see equation 4-2).

The individuals' choices are then aggregated to calculate the market share of each alternative path. The aggregate market share is the decision variable that is actually useful to transit agencies. Suppose the population size T is known then the total expected number of individuals choosing alternative i, denoted by $N_T(i)$, would simply be,

$$N_T(i) = \sum_{n=1}^{T} P(i \mid X_n) \qquad (4\text{-}4)$$

in which X_n are the attributes used in the utility function (Ben-Akiva 1985).

In this example, the population size is 492 trips from Quincy to Paulina. Because there are only trip-specific attributes and no traveler-specific attributes in the model specification, the probabilities of choosing alternative i are the same across individuals. The equation can be simplified to $N_T(i) = P(i \mid X_n) \times T$. In another word, the probability of a person choosing each alternative is numerically equal to the market share of that alternative in this particular model specification.

[12] Model E is a mixed logit model. A simulation is needed to accurately apply the model. Here I apply the model in its simplified form.

The predicted market shares for both paths for journeys between Quincy and Paulina are compared with the observed ones in Table 4-20. There is a 4.3% over prediction for SB path and 7.4% under prediction for NB path.

Table 4-20 Predicted versus Observed Market Shares of NB and SB Paths from Quincy to Paulina

Orig	DEST	Path	DIR	Observed Trips	Predicted Trips	Error Percentage
Quincy	Paulina	1	SB	311	324	4.3%
Quincy	Paulina	2	NB	181	168	-7.4%

The in-vehicle travel time and number of transfers are chosen as policy variables to be adjusted in the proposed network change. The predicted market shares are calculated based on the adjusted values of in-vehicle travel time and number of transfers. Figure 4-8 illustrates the impact of the in-vehicle travel time changes in the north bound path on the market share of each alternative. The right vertical line indicates the current level of IVT (18.8 minutes) and the white circles point to the corresponding predicted market shares: 34% on the NB path and 66% on the SB path. Suppose that the IVT on the north bound is decreased by 1 minute, other LOS attributes being constant, the black circles indicate the predicted market shares: 49% on the NB path and 51% on the SB path. The market shares are very sensitive to the changes in in-vehicle travel time.

Figure 4-8 Impact of NB Path In-vehicle Travel Time on the Market Shares of the NB and SB Paths

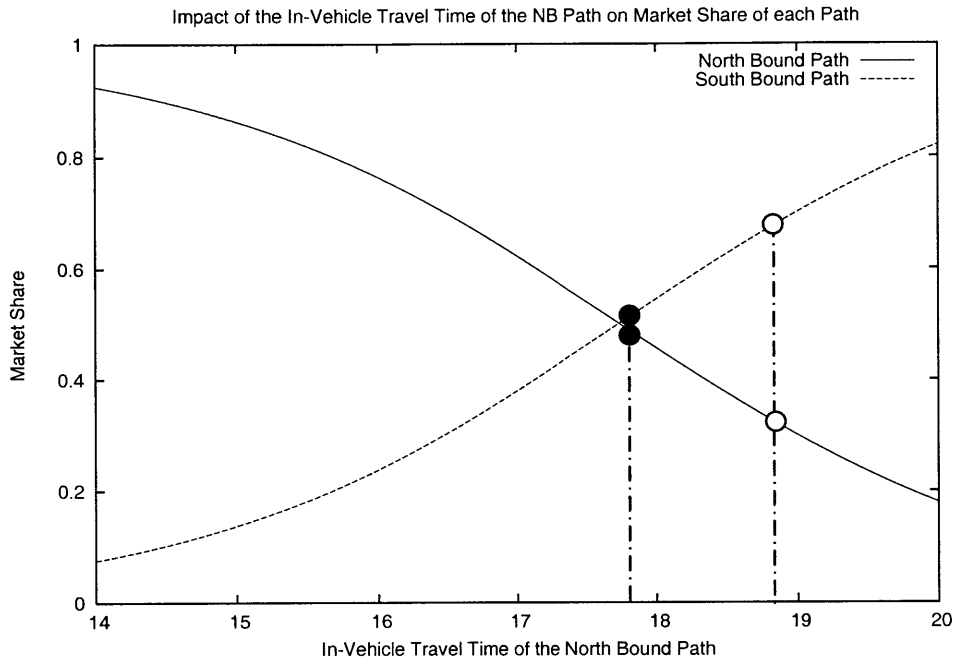
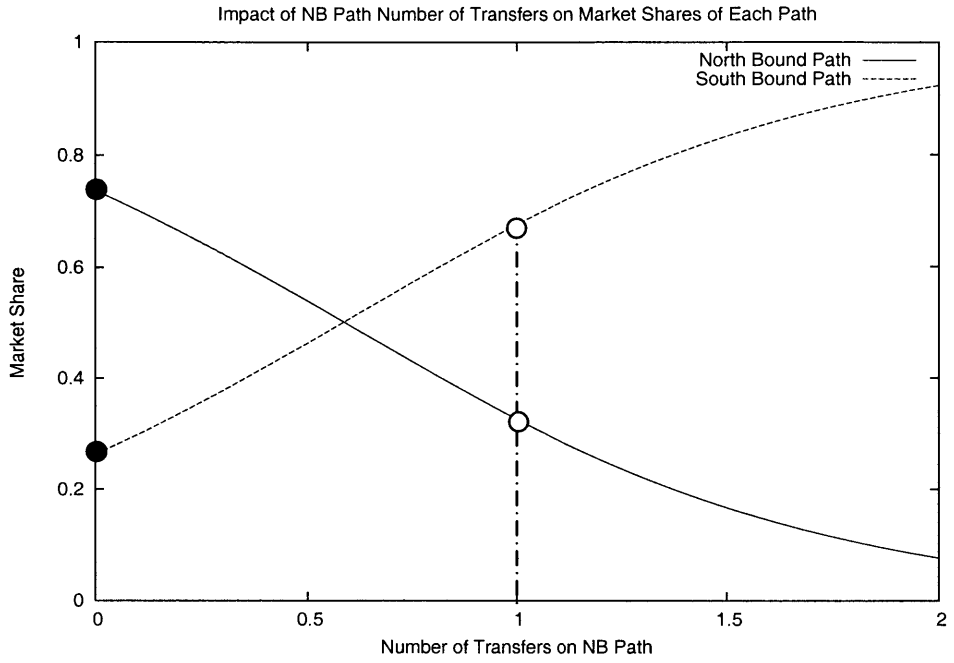


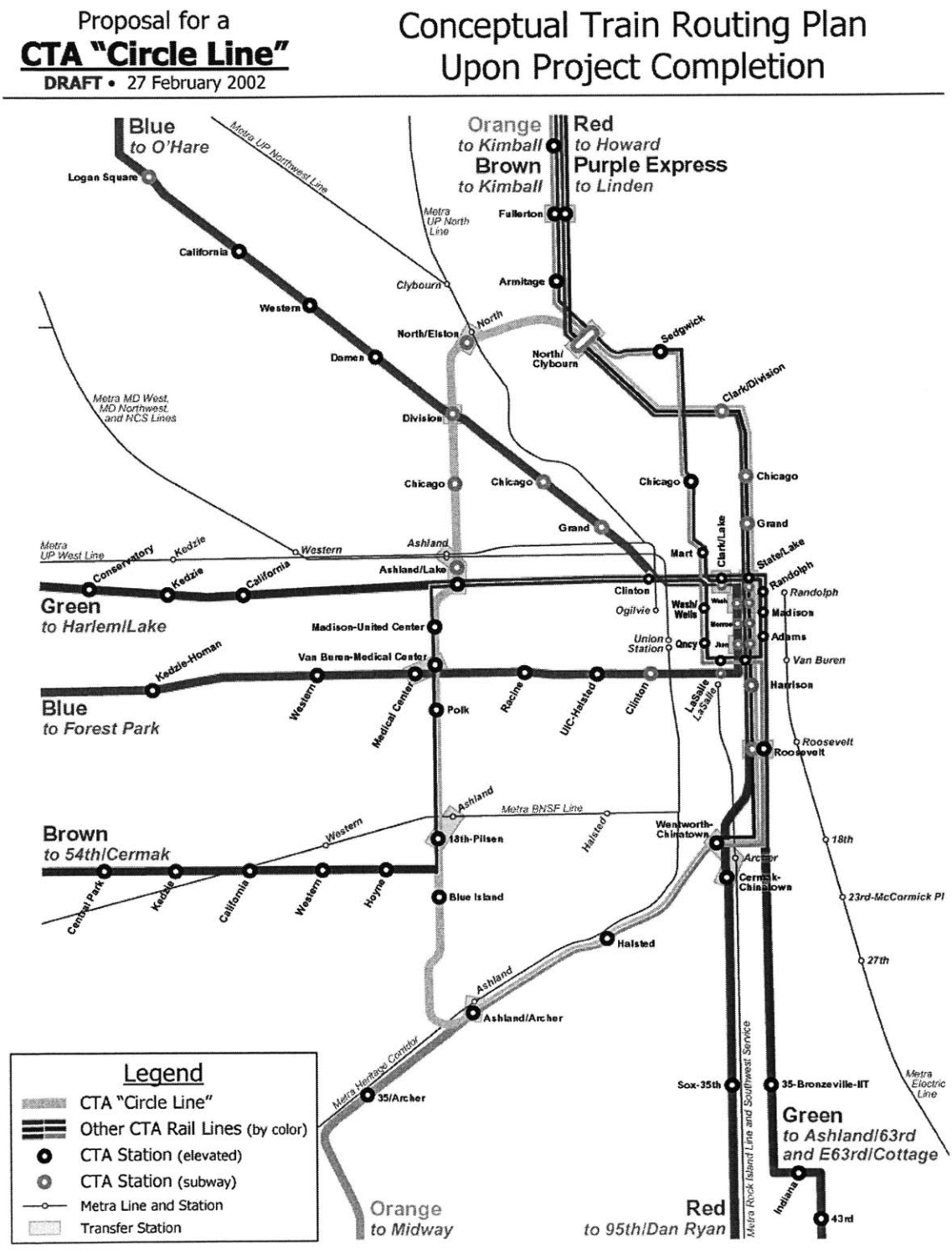
Figure 4-9 illustrates the impact of the number of transfers on the north bound path on the market share of each alternative. The x-axis shows the number of transfers on the north bound path. All other attributes of the north bound path and all the attributes of the south bound path are unchanged. While the number of transfers should obviously be a discrete value, the figure shows it as a continuous variable to clarify the presentation. The current market shares are indicated by the white circles. Suppose that in a new service plan, there is no transfer needed on the NB path, other LOS attributes being constant, the predicted market shares will change dramatically as indicated by the black circles: 75% on the NB path and 25% on the SB path.

Figure 4-9 Market Shares of NB and SB Paths with Changes in NB Path Number of Transfers



This example demonstrates that the model can be used to predict the market share changes among alternative paths resulting from changes of one (or more) LOS attributes. This case shows the process for only a single OD pair but this process can be repeated for all OD pairs in the network and the market shares of alternative paths of all OD pairs can be aggregated to obtain the demand on each service line. The CTA's circle line plan together with the rerouting of the existing service lines will change the path attributes for many OD pairs dramatically. Figure 4-10 illustrates the conceptual train routing plan of the CTA Circle Line project. This model can be applied to assess the impact of these changes upon the market shares of alternative paths, and then through aggregation, the demand for each service line.

Figure 4-10 Conceptual Train Routing Plan for CTA Circle Line Project



Chapter Five: Conclusion

Many ADC systems in public transit agencies have been designed and implemented with a single-minded goal: to announce bus stop information, to collect fares, to report emergencies, etc. This thesis proposes another perspective—if each of these systems is looked at as an individual data collection technology, a brand new method of data collection arises. Once the ADC systems are installed, they continuously generate huge volumes of raw data. This thesis examines the benefits as well as the difficulties of utilizing data from these ADC systems, and demonstrates that at very low marginal cost, the ADC systems can produce a very rich information base to support decision-making in public transit agencies. The three ADC data processing examples and two in-depth case studies in this thesis are preliminary efforts to set up a robust framework including data manipulation and analysis methodologies and techniques for the effective utilization of ADC systems. The utilization of ADC systems opens far more research avenues than are covered in this thesis, which is just the tip of the iceberg of possibilities.

Section 5.1 summarizes the research findings; Section 5.2 points out the limitations and difficulties in the current studies; and Section 5.3 proposes future research directions.

5.1 Overview of research findings

5.1.1 Enhanced Rail OD Matrix Inference Method

Traveler OD matrices are fundamental information required in public transit planning and operations analysis. This research proposed an enhanced method of inferring the rail trip OD matrix from an origin-only AFC system as an alternative to traditional passenger surveys. In the CTA rail system application, the enhanced method successfully inferred the destinations for 65% of all rail trip segments, 85% of which were inferred by the "TT" procedure employed by the prior studies and 15% were inferred by the "TB" procedure and trip chain symmetry utilization proposed in this research.

The proposed algorithm takes advantage of the pattern of a person's consecutive transit trip segments and uses the next trip segment boarding location information to infer the destination of the prior trip segment. There are two different cases in the usage of the next trip segment information: rail-to-rail and rail-to-bus. Prior studies only examined the rail-to-rail case and ignored the rail-to-bus case. This research proposed three strategies to examine the rail-to-bus case given different system availabilities. 1) When neither AVL data nor GIS technology are available to the transit system, the bus trip segment information cannot be utilized. 2) When GIS technology is employed but not AVL, the bus trip segment information can be partially utilized by examining the spatial connectivity between bus routes and the rail network. 3) When both GIS technology and the AVL data are available, the AFC data and the AVL data can be integrated to fully utilize the bus trip segment information in destination inference. In the CTA application,

the introduction of the GIS technology and the integration of AFC and AVL data account for 11% of the inferred destinations.

The daily and weekly trip chain patterns are observed and examined. The symmetry in some trip chains is utilized to improve the destination inference algorithm. The algorithm explicitly checks whether the bus routes of the two B trips are the same to verify the symmetry. When symmetry is found, the later T origin is used to infer the earlier T destination so that the possibility of the inference is expanded from only consecutive trip segments to some non-consecutive but symmetrical trip segments. This expansion helps to infer destinations of 2.4 percent more trips.

The implementation of the algorithm was initially cumbersome: the processing of one day's AFC and AVL data and the destination inference took about a month to finish when the algorithm was first proposed. A software tool was developed to speed up this process. This tool takes the raw AFC and AVL data and several accessory tables as input and calculates the rail OD matrix automatically. This tool does not require the users to have a detailed knowledge of the ADC systems or a deep understanding of database and GIS technologies. The software tool was applied to the CTA rail system for a one-week period from Jan.11th-17th, 2004. The AFC and the AVL data from CTA were input and the one week rail trip OD matrix was produced in about one hour. This software enables the inference algorithm to be practical in public transit agencies.

5.1.2 Rail Path Choice Modeling

Many rail travelers have alternative paths available for some journeys and they make choices over these alternatives. Discrete choice modeling is utilized to characterize how travelers make such decisions. This study is based on the CTA rail system and takes advantage of two data sources: the inferred rail trip OD matrix and the attributes of alternative paths calculated from a network representation in Trans CAD. This case study demonstrates that a rigorous travel behavior analysis can be performed based on data from the ADC systems.

In-vehicle travel time, number of transfers and transfer walking time are found to be the most important factors people consider in their path choice process. The waiting times (both the initial waiting time and the transfer waiting time) are not significant in the CTA rail system where there is little variation in the waiting times for alternative paths. This is consistent with the findings in Guo's (2003) study of MBTA.

To most of the travelers, walking in a transfer station is perceived more negatively than traveling in a train. On average one minute of transfer walking time is perceived by a typical traveler as equivalent to more than three minutes of in-vehicle travel time.

There is a trade off between making a transfer and in-vehicle travel time savings. When only in-vehicle travel time and number of transfers are included in the model (Model A), one transfer is perceived as equivalent to 9 minutes of in-vehicle travel time. This can be interpreted as the total transfer penalty measured in in-vehicle travel time. This value is

very close to the 10 minutes/transfer, found in Guo's MBTA study in 2003. With more factors entering the model, this ratio decreases because the transfer penalty is partially accounted for by the new factors. In the model with transfer attributes included (Model B1), the transfer penalty is found to be 6.9 minutes of in-vehicle travel time.

In-vehicle travel time is perceived differently for trips of different lengths: the longer the trip, the less travelers care about the travel time difference between the alternatives. However, travelers' preference for fewer transfers is found to be independent of the trip length. Consequently, the transfer penalties calculated as the ratio of the coefficients between number of transfers and in-vehicle travel time differ by trip length. The longer the trip, the higher the transfer penalty is measured in equivalent in-vehicle travel time. In the full model (Model E), for the average length trip the transfer penalty is 2.8 minutes of in-vehicle travel time, while for the longest trip the transfer penalty becomes 5 minutes of in-vehicle travel time.

The mixed logit model with random coefficients was tested to examine travelers' taste variation over different LOS variables. The results suggest that:

1) there is significant taste variation among travelers with respect to the transfer walking time. If assumed normal, the distribution of travelers' different perceptions can be characterized as a normal distribution with a mean of -2.4 and standard deviation of 1.9;
2) there is no significant variation in travelers' taste over the number of transfers. Together with the finding that travelers' preference for fewer transfers is

independent of the trip length, it is concluded that the transfers are perceived very similarly by different travelers regardless of trip length; and

3) controlling for the trip length and the number of stops, travelers' preferences for in-vehicle travel time are similar. In other words, travelers' taste variations with respect to in-vehicle travel time can be explicitly captured by the trip length and the number of stops.

Due to limitations of the data set, there is no traveler-specific information in the model: all variables are trip-specific. Nonetheless, the model still has good explanatory power, indicated by an adjusted ρ^2 of 0.37.

The application of the model in the CTA context with a proposed change in network attributes demonstrates that the market shares of alternative paths are quite sensitive to changes in the network attributes. In one example for journeys between Quincy and Paulina, a six-minute change in the in-vehicle travel time in the northbound path, all else being equal, could completely change the distribution of travel flows along the alternative paths—from the northbound path carrying 90% of these trips to it carrying only 20% of these trips. This model is shown to be useful in predicting the market shares of alternative paths under proposed network configuration changes. The market shares of alternative paths between all the OD pairs can be aggregated to obtain the spatial distribution of the total demand on the rail network. The service frequency and capacity for each service line can be planned accordingly.

5.2 Limitations and Challenges

5.2.1 Rail OD Inference Algorithm Validation

There are four issues that affect the coverage and the accuracy of the rail OD matrix inference algorithm:

1) the AFC system does not record all the rail trips since the AFC card is not the only way that passengers can pay for their trips. Cash and paper transfer can also be used and while cash payment has become very rare, paper transfers from the previous bus trip segment still account for a sizable portion of rail trips. The percentage of customers using AFC cards varies by the time of day and by station, averaging 91%. Whether or not the 91% coverage accurately represents the travel behavior of the entire population requires further validation;
2) the three assumptions of the destination inference algorithm are yet to be fully verified although CTA has recently contracted a consulting firm to test the statistical validity of the destination inference algorithm.
3) the rail OD inference algorithm can infer the destinations for 65% of the recorded rail trips but this may not represent a random sample of all rail trips. There may be bias in the coverage rate of certain OD pairs; and
4) the January 2004 effective coverage of the CTA AVL system is quite low. Although about 70% of the buses had AVL equipment installed, not all these devices are running at all times. The effective coverage of CTA AVL (the ratio of the AVL running bus-hour over the total bus-hour) at the time of this analysis appears to be only 40%.

5.2.2 Limitations of path choice models

The ADC systems do not collect demographic data on travelers and there is currently no way to link the ADC data with other passenger surveys[13]. So in the path choice models, only trip-specific variables are included in the utility function specifications. Although the model results indicate a good data fit and high explanatory power, it would be better to include some travelers' socio-economic information in the utility function. The mixed logit model is utilized partially to reflect travelers' unobserved taste variations, however, with socio-economic data, much of this taste variation might be modeled explicitly.

5.3 Future Research Directions

There are three major extensions of the current research: 1) to expand the study from the rail system to the entire public transit system; 2) to apply the path choice model to the CTA's proposed Circle Line service plan; and 3) to study transit agencies with different challenges in ADC utilization.

5.3.1 Extension from rail to the full public transit system and the linkage with land use and demographic characteristics

The current study is focused on the rail system—only the path alternatives on the rail network are considered. However, with AVL data now becoming widely available, there is the potential to observe the whole public transit trip chain including both rail and bus

[13] This situation is gradually changing with the implementation of smart cards in CTA, which records the travelers' identification in the card. However there will certainly be restrictions on the use of this data.

by combining the AFC data and the AVL data with GIS and relational database technology.

The expansion from rail to the entire public transit system opens tremendous research possibilities. The path choice decision becomes much more complicated involving both mode choices and path choices simultaneously. When the origin and the destination of the trip can be pinpointed to the bus stop level, travel patterns can be connected to land use patterns around the trip origin and destination. Furthermore the average demographic information of the trip origin and destination areas from census statistics based on census tracts or even census blocks can be combined with the land use attributes and the trip-specific LOS attributes to more carefully examine travelers mode and path choice in the public transit system.

5.3.2 Application in the CTA Circle Line Plan Evaluation

The path choice model characterizes travelers' path choices in the rail system. The model can be used to forecast the probability that an individual chooses a certain rail path for an OD pair. Individuals' choices collectively determine the market shares of the alternative paths for the OD Pair. The spatial distribution of the total demand on the rail network can be determined by aggregating the market shares of alternative paths between all OD pairs. One important application is to the Chicago Transit Authority (CTA) rail system, where a new circle line and service line rerouting in the loop area have been proposed with the first stage slated for implementation in spring, 2005. This application of the model would facilitate the evaluation of the options of the circle line plan by examining how its implementation would change the spatial distribution of flow in the network.

5.3.3 Case studies of transit agencies with different ADC utilization challenges

This thesis takes CTA as an example to examine the characteristics of the ADC systems and demonstrates the potential benefits. There are many different types of the ADC systems implemented in public transit authorities in the U.S. and around the world. It would be enlightening to study other cases with different technological characteristics and different institutional challenges.

Take the automated fare collection system (E-Z link) used in Singapore's public transit systems, for example. Technically Singapore's E-Z link has many advantages over the CTA's AFC: 1) EZ link covers both rail and bus. Most people use E-Z Link card to pay for their trips; 2) EZ link has both entry and exit control for both rail and bus trips because the fare is distance related; 3) When EZ link is used on buses, the boarding and alighting locations are recorded; and 4) The EZ link card has a much longer life with many travelers using the same card for months. However there is a big institutional obstacle to access and utilize the E-Z link data. The E-Z link data are not supposed to be public. The Land Transportation Authority in Singapore supervises the public transit system but does not directly run any rail or bus services. All the rail and bus services are operated by private companies such as SMRT Inc. and SBS Inc., which have a major concern about releasing any data to actual or potential competitors. So how to utilize the data collected by E-Z link without infringing on the business prospects of the competing bus/rail operators poses quite a different challenge than the ones facing the CTA.

5.3.4 Institutional responses and system design initiative

There are many opportunities to improve planning and forecasting in public transit agencies using technologies developed in this thesis, especially true for large metro areas with well developed systems with multiple path choices—where analysis is too complex and computationally difficult for traditional methods. As ITS unfolds, it will be increasingly important to take advantage of the new algorithm and path choice models.

However, many questions need to be answered before the findings of this research can be applied in transit agencies: what should be the institutional responses from the transit agencies? How should ADC systems data be coordinated with traditional data surveys? For transit agencies which are about to implement ADC systems, what are the lessons and experiences they should learn? What are the criteria for selecting the right vendor of ADC systems? How can we encourage different vendors to coordinate with each other so that the data from multiple systems can be seamlessly integrated?

Appendix One: C/C++ Code for Pre-Destination Inference Data Processing

/* The C/C++ source code is stored as afcPRE.c. It is complied on Athena with GNU GCC complier. */

```
/* Contants Definition */
#define sizeAFC 5948454 /* 5948454 total AFC record between Jan11 to Jan17 */
#define sizeAFCIndex 1076867  /* 1076867  distinct sn in afc table */
#define sizeAFCBT 2657680 /* 2657680 total number of trip with BusID infered ????*/
#define sizeAFCBTIndex 1724 /* 1724 distinct busid recorded in AFC table ????*/
#define sizeAVL 1618009 /* 1618009 total AVL record between Jan11 to Jan17 */
#define sizeAVLIndex 1219 /*1219 distinct busid in avl table */
#define sizeR2P 54 /* unique rail to pace transfer table size */
#define sizeE2BU 1898 /* unique equip to bus */
#define sizeE2BM 445 /* multiple equip to bus */
#define sizeS2S 12458 /* stop to station distance all 12458 records, one for each bus stop
*/
#define sizeS2S5 2170 /* 5 min stop to station transfer, 2170 stops are inside the distance
<=1320 feet */
#define sizeE2E 962 /* equip to entrance */
#define sizeA2S 153 /* AFC LOC ID to station */
#define sizeE2S 201 /* entrance to station */
#define MAX_TIME_SECOND 604800  /* 3600*24*7days */
#define MIN_TIME_SECOND 0
/* Type Definitions Omitted */
/* Comparison Function Definition Omitted */
int main(void) {
/* Variable Definition Omitted*/
/* Memory allocation Omitted */
/* Data Input IO operation Omitted */
/* Sort AFC table by SN then by datetime */
        qsort(afc, sizeAFC, sizeof(TypeAFC), compareAFC);
/* Create index on AVL busid Omitted */
/* Create index on AFC sn Omitted */
/* Update AFC 1st round*/
        for (i = 0; i < sizeAFC; i++) {
                if (afc[i].event == 36|| afc[i].event == 37 || afc[i].event == 38 || afc[i].event
== 39 || afc[i].event == 43 || afc[i].event == 45 || afc[i].event == 46 || afc[i].event == 47 ||
afc[i].event == 48 || afc[i].event == 49 || afc[i].event == 97 || afc[i].event == 98 ||
afc[i].event == 99) afc[i].ra = 1; else afc[i].ra = 0; /*to obtain RA */
                if (afc[i].curroute > 1024) afc[i].tb = 1; else afc[i].tb = 0; /* to obtain TB
*/
                /* to obtain LPACE LPACE ID */
```

```
            if (afc[i].preroute > 2048) { afc[i].lpace = 1; afc[i].lpaceid =
        afc[i].preroute - 2048; } else  afc[i].lpace = 0;
            if (afc[i].lpace == 1) { /* to obtain LPACETFR */
                short tmplpaceid = afc[i].lpaceid;
                for (j = 0; j < sizeR2P; j++) if (tmplpaceid == r2p[j].lpaceid)
{afc[i].lpacetfr = r2p[j].stationid; break;}}
            if (afc[i].tb == 0) {     /* bus trip */
                afc[i].busrouteid=afc[i].curroute; /* BUS Route ID */
                /* to obtain BUS ID*/
                pointerE2BU=(TypeE2BU *) bsearch(&afc[i].equipid, e2bu,
sizeE2BU, sizeof(TypeE2BU), compareE2BU);
                if (pointerE2BU!=NULL)      afc[i].busid = (*pointerE2BU).busid;
                else {
                    for (j = 0; j < sizeE2BM; j++) {
                    if (strcmp(afc[i].equipid, e2bm[j].equipid) == 0 &&
afc[i].datetime >= e2bm[j].min
                        && afc[i].datetime <= e2bm[j].max) {afc[i].busid =
e2bm[j].busid; break;}
                    }
                }
            } else /*tb=1, train trip*/ {
                afc[i].afclocid = afc[i].curroute - 1024; /* to obtain AFCLOCID
*/
                pointerA2S=(TypeA2S *) bsearch(&afc[i].afclocid, a2s, sizeA2S,
sizeof(TypeA2S), compareA2S);
                if (pointerA2S!=NULL) afc[i].station1 = (*pointerA2S).stationid;
        /* to obtain station1*/
                for (j = 0; j < sizeE2E; j++) /* to obtain entranceid */
                    if (strcmp(afc[i].equipid, e2e[j].equipid) == 0)
{afc[i].entranceid = e2e[j].entranceid; break;}
                if (afc[i].entranceid != 0) { /* to obtain station2 */
                    pointerE2S= (TypeE2S*) bsearch (&afc[i].entranceid, e2s,
sizeE2S, sizeof(TypeE2S), compareE2S);
                    if (pointerE2S!=NULL) afc[i].station2 =
(*pointerE2S).stationid;
                }
                /* to obtain Stationid from Station1 or 2 */
                if (afc[i].station1==afc[i].station2) afc[i].stationid=afc[i].station1;
                else if (afc[i].station1==9999) afc[i].stationid=afc[i].station2;
                else if (afc[i].station2==0) afc[i].stationid=afc[i].station1;
            }
        }
/* Update AFC 2nd round: combining AVL and AFC to get systemstopid*/
        for (i = 0; i < sizeAFC; i++) {
            if (afc[i].busid != 0) {
                pointer = -1;
```

```
                        tmptime = MIN_TIME_SECOND;
                        for (j = 0; j < sizeAVLIndex; j++) if (afc[i].busid ==
avlindex[j].busid) {pointer = j; break; }
                        if (pointer == -1) continue;
                        for (j = avlindex[pointer].position; j < avlindex[pointer].position +
avlindex[pointer].length; j++)
                                if (afc[i].datetime > avl[j].datetime && afc[i].datetime <
avl[j].datetime + 300)
                                        if (tmptime < avl[j].datetime) {tmptime =
avl[j].datetime;
                                                afc[i].systemstopid = avl[j].systemstopid;}
                }
        }
/* Update AFC 3rd round: calculate TFR5min*/
        for (i = 0; i < sizeAFC; i++) {
                if (afc[i].systemstopid != 0) {
                        pointerS2S=(TypeS2S *) bsearch(&afc[i].systemstopid, s2s5,
sizeS2S5, sizeof(TypeS2S), compareS2S);
                        if (pointerS2S!=NULL) afc[i].tfr5min = (*pointerS2S).stationid;
                }
        }
/* Result File Export Omitted */
/* Other Cleanup Omitted */
}
```

Appendix Two: C/C++ Code for Destination Inference Algorithm

```
/* The C/C++ source code is stored as afcDI.c. It is complied on Athena with GNU GCC complier. */
/* Contants Definition */
#define sizeAFC 5948454 /* 5948454*/
#define sizeAFCIndex 1076867 /* 1076867 distinct sn in afc table */
#define sizeBR2SU 39   /* 39 busroutes have Unique BusRoute2Station tfr */
#define MAX_TIME_SECOND 604800  /* 3600*24*7days */
#define MIN_TIME_SECOND 0
#define numDay 7 /* number of days in the time period, 7 for Jan11 to Jan17*/
#define MSTimeGap 180 /* 3minutes */
/* Inference Methods Definition */
#define UniqueLPace 1
#define DirectBus 2
#define UniqueBus 3
#define NextTrain 4
#define LastTrain 5
#define Special 6
#define SpecialB 7
#define MSUniqueLPace 11
#define MSDirectBus 12
#define MSUniqueBus 13
#define MSNextTrain 14
#define MSLastTrain 15
#define NoSolution 9999   /* used both in "method" and in "dest" */
#define MSNoSolution 19999
/* Omitted: Type Definition*/
/* Omitted: Compare Function Definition*/
int main(void) {
/* Omitted: Variables Definition*/
/* Omitted: Memory allocation steps*/
/* Omitted: Data Input */
/* Create index on sn of AFC Step 1*/
        k = 0;
        afcindex[k].sn = afc[0].sn; afcindex[k].position[0] = 0; afcindex[k].wlength = 1;
        for(i=0;i<numDay;i++) afcindex[k].length[i]=0;
        if (afc[k].datetime<10800) day=0;  else day= (afc[k].datetime-10800)/86400; /*
day boundary at 3am */
        afcindex[k].length[day]++;
        for (i = 0; i < sizeAFC - 1; i++) {
                if (afc[i+1].datetime<10800) day=0;
                else day= (afc[i+1].datetime-10800)/86400;
                if (afc[i].sn != afc[i + 1].sn) {
```

```
                        k++; afcindex[k].sn = afc[i + 1].sn; afcindex[k].wlength = 1;
afcindex[k].position[0] = i + 1;
                        for(j=0;j<numDay;j++) afcindex[k].length[j]=0;
                        afcindex[k].length[day] = 1;
                } else  { afcindex[k].wlength++; afcindex[k].length[day]++; }
        }
        printf("\n AFC Indexing Step 1 Complete\n Total %d index records \n", k + 1);
/* Create index on sn of AFC Step2: set position for each day */
        for (i=0;i<sizeAFCIndex;i++)
                for(day=0;day<numDay-1;day++)
        afcindex[i].position[day+1]=afcindex[i].position[day]+afcindex[i].length[day];
        printf("AFC Indexing Step 2 Complete \n");
/* Set chain, chain2, train */
for (i=0;i<sizeAFCIndex;i++) {
        for(day=0;day<numDay;day++) {
                /* allocate memory for char * chain and char * chain2 */
                afcindex[i].chain[day]= (char *)
calloc(afcindex[i].length[day]>0?afcindex[i].length[day]:1, sizeof(char));
                if (afcindex[i].chain[day] == NULL) { printf("\nRun out of memory! \n");
exit(1);  }
                afcindex[i].chain2[day]= (char *)
calloc(afcindex[i].length[day]>0?afcindex[i].length[day]:1, sizeof(char));
                if (afcindex[i].chain[day] == NULL) { printf("\nRun out of memory! \n");
exit(1); }
                afcindex[i].train[day]=0; afcindex[i].bus[day]=0;
        for(j=afcindex[i].position[day];j<afcindex[i].position[day]+afcindex[i].length[day
];j++)
                if (afc[j].ra==1) {
                        if (afc[j].tb==1) {
                                afcindex[i].chain[day]=strcat(afcindex[i].chain[day],"T");
                                afcindex[i].chain2[day]=strcat(afcindex[i].chain2[day],"T");
                                afcindex[i].train[day]++;
                        } else {
                                afcindex[i].chain[day]=strcat(afcindex[i].chain[day],"B");
                                afcindex[i].chain2[day]=strcat(afcindex[i].chain2[day],"B");
                                afcindex[i].bus[day]++;
                        }
                } else  afcindex[i].chain2[day]=strcat(afcindex[i].chain2[day],"A");
                if (strcmp(afcindex[i].chain2[day],"") == 0 ) afcindex[i].chain2[day]="0";
                if (strcmp(afcindex[i].chain[day],"") == 0 ) afcindex[i].chain[day]="0";
        }
}
                printf("AFC Indexing Step 3 Complete \n");
/* Export trip chain pattern*/
        ftripchain= fopen("/mit/duspbucket/nTripChain1.txt", "w");
```

```
        if (ftripchain == NULL) { printf("The file /mit/duspbucket/nTripChain1.txt can
not be opened!\n"); exit(0); }
        else printf("\n File /mit/duspbucket/nTripChain1.txt Export Start!\n");
        for (i=0;i<sizeofAFCIndex;i++)
        {       fprintf(ftripchain, "%d,%d",afcindex[i].sn,afcindex[i].wlength);
                for (j=0;j<numDay;j++) fprintf(ftripchain, ",%s", afcindex[i].chain[j]);
                for (j=0;j<numDay;j++) fprintf(ftripchain, ",%s", afcindex[i].chain2[j]);
                fprintf(ftripchain, "\n"); }
        fclose(ftripchain);
```

/*Destination inference core algorithm */

```
for (i=0;i<sizeAFCIndex;i++)
for (day=0;day<numDay;day++) {
MSCount=0;
for(j=afcindex[i].position[day];j<afcindex[i].position[day]+afcindex[i].length[day];j++) {
if (afc[j].ra==1 && afc[j].tb==1 && afc[j].stationid!=0) {          /* only consider
actual train trips*/
        MSwipe[MSCount]=j;
        MSCount++;
        if (j!=afcindex[i].position[day]+afcindex[i].length[day]-1) {
                k=j+1;
                if(afc[k].lpaceid!=0 ) {                    /* if previous trip is pace bus*/
                        if (afc[k].lpacetfr!=0 && afc[k].lpacetfr!=afc[j].stationid) {
                        /* if previous trip is pace bus and the tfr station is reasonable*/
                                for(n=0;n<MSCount;n++) {
                                        afc[MSwipe[n]].dest=afc[k].lpacetfr;
                                        afc[MSwipe[n]].method=(n==MSCount-
1?UniqueLPace:MSUniqueLPace);
                                }
                                MSCount=0;
                        } else{         /* if previous trip is pace bus and the tfr station is
NOT reasonable, no solution for trip j*/
                                for(n=0;n<MSCount;n++) {
                                        afc[MSwipe[n]].dest=NoSolution;
                                        afc[MSwipe[n]].method=(n==MSCount-
1?NoSolution:MSNoSolution);
                                }
                                MSCount=0;
                        }
                } else if (afc[k].tb==1) {                  /* next is train */
                        if (afc[j].stationid!=afc[k].stationid && afc[k].stationid!=0){
                        /* orig <> dest AND k.orig has a stationid, reasonable trip */
                                for(n=0;n<MSCount;n++) {
                                        afc[MSwipe[n]].dest=afc[k].stationid;;
                                        afc[MSwipe[n]].method=(n==MSCount-
1?NextTrain:MSNextTrain);
```

```
                        }
                        MSCount=0;
                    } else if((afc[k].datetime-afc[j].datetime>MSTimeGap) ||
afc[k].ra!=1 || afc[k].tb!=1 || afc[k].stationid==0) {
                        /* orig==dest but Not multiple swipe */
                            for(n=0;n<MSCount;n++) {
                                afc[MSwipe[n]].dest=NoSolution;
                                afc[MSwipe[n]].method=(n==MSCount-
1?NoSolution:MSNoSolution);
                            }
                            MSCount=0;
                    }
                } else if (afc[k].tb==0 && afc[k].ra==1) { /* next is bus*/
                    if (afc[k].tfr5min!=0 && afc[k].tfr5min!=afc[j].stationid)   {/* if
direct train to bus tfr reasonable */
                        for(n=0;n<MSCount;n++) {
                            afc[MSwipe[n]].dest=afc[k].tfr5min;
                            afc[MSwipe[n]].method=(n==MSCount-
1?DirectBus:MSDirectBus);
                        }
                        MSCount=0;
                    } else  {               /* if no direct train to bus tfr then check if
there is bus route to train unique transfer*/
                        pointerBR2SU=(TypeBR2SU *)
bsearch(&afc[k].busrouteid, br2su, sizeBR2SU, sizeof(TypeBR2SU), compareBR2SU);
                        if (pointerBR2SU!=NULL)   {
                            for(n=0;n<MSCount;n++) {

      afc[MSwipe[n]].dest=(*pointerBR2SU).stationid;
                                afc[MSwipe[n]].method=(n==MSCount-
1?UniqueBus:MSUniqueBus);
                            }
                            MSCount=0;
                        } else {
                            for(n=0;n<MSCount;n++) {
                                afc[MSwipe[n]].dest=NoSolution;
                                afc[MSwipe[n]].method=(n==MSCount-
1?NoSolution:MSNoSolution);
                            }
                            MSCount=0;
                        }
                    }
                } else {
                        for(n=0;n<MSCount;n++) {
                            afc[MSwipe[n]].dest=NoSolution;
```

```
                                        afc[MSwipe[n]].method=(n==MSCount-
1?NoSolution:MSNoSolution);
                                }
                                MSCount=0;
                        }
                } else {
                        if(afc[j].stationid!=afc[afcindex[i].position[day]].stationid &&
afc[afcindex[i].position[day]].stationid!=0)  {
                                for(n=0;n<MSCount;n++) {   /* assign j.dest =
firstTrain.orig */

                afc[MSwipe[n]].dest=afc[afcindex[i].position[day]].stationid;
                                        afc[MSwipe[n]].method=(n==MSCount-
1?LastTrain:MSLastTrain);
                                }
                                MSCount=0;
                        } else {
                                for(n=0;n<MSCount;n++) {
                                        afc[MSwipe[n]].dest=NoSolution;
                                        afc[MSwipe[n]].method=(n==MSCount-
1?NoSolution:MSNoSolution);
                                }
                                MSCount=0;
                        }
                }
}}}
printf("\nDestination inferenece algorithm finished!\n");
printf("\nDestination inferenece Special Cases Start!\n");

/* Special Trip Chain Consideration*/
for (i=0;i<sizeAFCIndex;i++)
        for (day=0;day<numDay;day++) {

/*TBBT*//*TBBTB*//*TBBTT*//*TBBTTB*//*TBBTBB*//*TBBTTT*/
                if (strcmp(afcindex[i].chain[day],"TBBT")==0 ||
strcmp(afcindex[i].chain[day],"TBBTB")==0
||strcmp(afcindex[i].chain[day],"TBBTT")==0  ||
strcmp(afcindex[i].chain[day],"TBBTTB")==0
||strcmp(afcindex[i].chain[day],"TBBTBB")==0 ||
strcmp(afcindex[i].chain[day],"TBBTTT")==0)
                {       j=afcindex[i].position[day];
                        if (afc[j].dest==NoSolution &&
afc[j+1].busrouteid==afc[j+2].busrouteid
                                && afc[j+3].lpaceid==0 && afc[j+3].stationid!=0 &&
afc[j+3].stationid!=afc[j].stationid) {
```

```
                                        afc[j].dest=afc[j+3].stationid;
        afc[j].method=Special;}
                }
/*BTBBT*//*TTBBT*//*BTBBTBB*//*TTBBTT*//*BTBBTTB*/
                else if (strcmp(afcindex[i].chain[day],"BTBBT")==0 ||
strcmp(afcindex[i].chain[day],"TTBBT")==0 ||
strcmp(afcindex[i].chain[day],"BTBBTBB")==0
||strcmp(afcindex[i].chain[day],"TTBBTT")==0
||strcmp(afcindex[i].chain[day],"BTBBTTB")==0)
                {       j=afcindex[i].position[day]+1;
                        if (afc[j].dest==NoSolution &&
afc[j+1].busrouteid==afc[j+2].busrouteid
                                && afc[j+3].lpaceid==0 && afc[j+3].stationid!=0 &&
afc[j+3].stationid!=afc[j].stationid) {
                                        afc[j].dest=afc[j+3].stationid;
afc[j].method=Special; }
                }
                else if (strcmp(afcindex[i].chain[day],"TBBBBT")==0) { /*TBBBBT*/
                        j=afcindex[i].position[day];
                        if (afc[j].dest==NoSolution &&
afc[j+1].busrouteid==afc[j+4].busrouteid && afc[j+2].busrouteid==afc[j+3].busrouteid
&& afc[j+4].lpaceid==0 && afc[j+4].stationid!=0 && afc[j+4].stationid!=afc[j].stationid)
                        { afc[j].dest=afc[j+4].stationid; afc[j].method=Special; }
                }
/*BTTBBTB*//*BBTBBTB*//*BTTBBT*/
                else if (strcmp(afcindex[i].chain[day],"BTTBBTB")==0 ||
strcmp(afcindex[i].chain[day],"BBTBBTB")==0 ||
strcmp(afcindex[i].chain[day],"BTTBBT")==0)
                {       j=afcindex[i].position[day]+2;
                        if (afc[j].dest==NoSolution &&
afc[j+1].busrouteid==afc[j+2].busrouteid  && afc[j+3].lpaceid==0 &&
afc[j+3].stationid!=0 && afc[j+3].stationid!=afc[j].stationid)
                        { afc[j].dest=afc[j+3].stationid; afc[j].method=Special; }
                }
/*B??????B*//*BTTB,BTBBTB,BTBTB,BTTTB,BTTTTB,BBTTBB,BTBBBTB,BTBT
TB,BTTBTB,BTBBTTB,BTTBBTB,BTTBBTTB*/
                        j=afcindex[i].position[day];
                        k=afcindex[i].position[day]+afcindex[i].length[day]-1;
                        if (k>j && afc[j].tb==0 && afc[k].tb==0 && afc[j+1].tb==1 &&
afc[k-1].tb==1 && afc[j].busrouteid ==afc[k].busrouteid && afc[k-1].dest==NoSolution
&& afc[j+1].stationid!=0 && afc[j+1].stationid!=afc[k-1].stationid)
                        { afc[k-1].dest=afc[j+1].stationid; afc[k-1].method=SpecialB; }

}
/*Omitted: Export Dest Result */
}
```

Appendix Three: Biogeme Model File for Mixed Logit Model E

```
// Mixed Logit Model E
// Mon Jun 28 05:52:47 2004
// Run by Jinhua Zhao
// BIOGEME Version 0.7 [Mon Dec 15 17:45:55 2003]
// Developed by Michel Bierlaire, EPFL (c) 2001-2003

[DataFile]
// Specify the number of columns that must be read in the data file
// It is used to check if the data file is read correctly.
$COLUMNS = 20

[Choice]
CHOICE

[Weight]
 WEIGHT   *  0.00974759

[Beta]
// Name        Value                 LowerBound          UpperBound          Status
(0=variable, 1=fixed)
BIVT         -8.6169494e-01 -1.0000000e+04 +1.0000000e+04 0
BIVTLEN      +1.7349917e-02 -1.0000000e+04 +1.0000000e+04 0
BIVTSTOPS    +7.5149087e-03 -1.0000000e+04 +1.0000000e+04 0
BIVT_S       +2.7184440e-01 +0.0000000e+00 +1.0000000e+04 0
BTFR         -1.7581292e+00 -1.0000000e+04 +1.0000000e+04 0
BTFR_S       +4.8017796e-02 +0.0000000e+00 +1.0000000e+04 0
BWALK        -2.4100657e+00 -1.0000000e+04 +1.0000000e+04 0
BWALK_S      +1.8876829e+00      +0.0000000e+00      +1.0000000e+04      0
[Mu]
// In general, the value of mu must be fixed to 1. For testing purposes, you
// may change its value or let it be estimated.
// Value LowerBound UpperBound Status
+1.0000000e+00       +0.0000000e+00       +1.0000000e+00       1

[SampleEnum]
// Number of simulated choices to include in the sample enumeration file
+0

[Utilities]
// Id Name  Avail  linear-in-parameter expression (beta1*x1 + beta2*x2 + ... )
```

```
1       Alt1    AVAL1       BIVT [ BIVT_S ] * IVT1 + BWALK [ BWALK_S ] *
WALK1 + BTFR [ BTFR_S ] * TFR1 + BIVTSTOPS * IVTSTOPS1 + BIVTLEN *
IVTLEN1
2       Alt2    AVAL2       BIVT [ BIVT_S ] * IVT2 + BWALK [ BWALK_S ] *
WALK2 + BTFR [ BTFR_S ] * TFR2 + BIVTSTOPS * IVTSTOPS2 + BIVTLEN *
IVTLEN2

[GeneralizedUtilities]
$NONE

[ParameterCovariances]
// Par_i Par_j Value  LowerBound UpperBound  status (0=variable, 1=fixed)
$NONE
[Expressions]
// Define here arithmetic expressions for name that are not directly
// available from the data
one  = 1
IVTSTOPS1  =  IVT1   *  STOPS1
IVTSTOPS2  =  IVT2   *  STOPS2
IVTLEN1  =  IVT1   *  LENGTH1
IVTLEN2  =  IVT2   *  LENGTH2
[Draws]
5000

[Group]
1

[Exclude]
0

[Model]
// Currently, the following models are available
// Uncomment exactly one of them
$MNL  // Multinomial Logit Model
//$NL  // Nested Logit Model
//$CNL  // Cross-Nested Logit Model
//$NGEV // Network GEV Model

[Scale]
// The sample can be divided in several groups of individuals. The
//utility of an individual in a group will be multiplied by the scale factor
//associated with the group.

// Group_number  scale LowerBound UpperBound status

1       +1.0000000e+00       +1.0000000e+00       +1.0000000e+00       1
```

```
[NLNests]
// Name paramvalue  LowerBound UpperBound  status list of alt
$NONE
[CNLNests]
// Name paramvalue LowerBound UpperBound  status
$NONE

[CNLAlpha]
// Alt Nest value LowerBound UpperBound  status
$NONE

[Ratios]
// List of ratios of estimated coefficients that must be produced in
// the output. The most typical is the value-of-time.
// Numerator   Denominator  Name
$NONE

[LinearConstraints]
$NONE

[NonLinearEqualityConstraints]
$NONE

[NonLinearInequalityConstraints]
// At this point, BIOGEME is not able to handle nonlinear inequality
// constraints yet. It should be available in a future version.
$NONE

[NetworkGEVNodes]
// All nodes of the Network GEV model, except the root,
// must be listed here, with their associated parameter.
// If the nodes corresponding to alternatives are not listed,
// the associated parameter is constrained to 1.0 by default
// Name  mu_param_value    LowerBound   UpperBound   status
$NONE

[NetworkGEVLinks]
// There is a line for each link of the network.
// The root node is denoted by __ROOT
// All other nodes must be either an alternative or a node listed in
// the section [NetworkGEVNodes]
// Note that an alternative cannot be the a-node of any link,
// and the root node cannot be the b-node of any link.
// a-node  b-node alpha_param_value LowerBound UpperBound  status
$NONE
```

Bibliography

Adam Rahbee and David Czerwinski, Using Entry-Only Automatic Fare Collection Data to Estimate Rail Transit Passenger Flows at CTA, Proceedings of the 2002 Transport Chicago Conference

Chicago Transit Authority Press Release 10/3/2002, Next stop announcements will make bus rides easier for visually or hearing impaired customers.

Chicago Transit Authority, Travel Behavior and Attitude Survey, 2001

Chicago Transit Authority, Request For Proposal: statistical validation of the farecard passenger flow model for federal NTD reporting, 2004

Erik Sheridan Wile, Use of Automatically Collected Data to Improve Transit Line Performance, Master of Science in Transportation thesis, MIT, September 2003

James J. Barry, Robert Newhouser and Rahbee, A., and Sayeda, S. Origin and Destination Estimation in New York City Using Automater Fare System Data. Proceedings of the 2001 TRB Planning Applications Conference, Corpus Christi, TX

Kenneth E. Train, Discrete Choice Methods with Simulation, Cambridge University Press, 2003

Michael Meyer and Eric Miller, Urban Transportation Planning a decision-oriented approach

Michel Bierlaire, An Introduction to Biogeme, 2003

Moshe Ben-Akiva and Michel Bierlaire, Discrete choice methods and their applications to short term travel decisions, Handbook of Transportation Science, edited by Randolph W. Hall

Moshe Ben-Akiva, Steven Lerman, Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, 1985, Cambridge, MA

Michael Scott Ramming, Network Knowledge and Route Choice, MIT PhD Dissertation, 2002

TransCAD, Transportation GIS Software, Travel Demand Modeling with TransCAD 4.0, Caliper, 2002

TransCAD, Transportation GIS Software, User Guide, Caliper, 2000

You-Lian Chu, 2002. Automobile Ownership Analysis Using Ordered Probit Models. Transportation Research Record, 1805.

Zhan Guo and Nigel H.M. Wilson, Assessment of the Transfer Penalty for Transit Trips: A GIS-based Disaggregate Modeling Approach, TRB, 2004

Zhan Guo, Assessment of the Transfer Penalty for Transit Trips in Downtown Boston--A GIS-based Disaggregate Modeling Approach, Master Thesis, MIT, 2003