Bus Passenger Origin-Destination Estimation and Travel Behavior Using Automated Data Collection Systems in London, UK

by

Wei Wang

B.S. in Transportation Engineering
Southeast University
Nanjing, China (2007)

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

Massachusetts Institute of Technology

June 2010

Author ....

Department of Civil and Environmental Engineering
May 24, 2010

Certified by..

Nigel H.M. Wilson
Professor of Civil and Environmental Engineering
Chair, Master of Science in Transportation Program
Thesis Supervisor

Certified by..

John P. Attanucci
Research Associate of Civil and Environmental Engineering
Thesis Supervisor

Accepted by..

Professor Daniele Veneziano
Chairman, Departmental Committee for Graduate Students

Bus Passenger Origin-Destination Estimation and Travel Behavior Using Automated Data Collection Systems in London, UK

by

Wei Wang

Submitted to the Department of Civil and Environmental Engineering
On May 24, 2010 in Partial Fulfillment of the
Requirements for the Degree of Master of Science in Transportation

ABSTRACT

This research explores the application of archived data from Automatic Data Collection Systems (ADCS) to transportation planning with a focus on bus passenger Origin-Destination (OD) inferences at the bus-route level and on travel behavior, using London as an example. This research demonstrates the feasibility and ease of applying the trip-chaining method to infer bus passengers' boarding and alighting locations, and validates the results by comparing them with the Bus Passenger Origin and Destination (BODS) survey data in London. With the inferred OD matrices, the variations of weekday and weekend bus route OD patterns over a two-week period are examined for planning purposes. Given these variations, reliance on ADCS can provide transit planners with more comprehensive, reliable and correct information for service planning than traditional manual surveys.

Moreover, while interchange conditions and performance are considered important inputs for public transit planning, collecting such data has not been easy. Based on the inferred OD matrices and the Automatic Vehicle Location (AVL) data, alighting times for bus passengers can also be estimated. As a result, bus journey stages can easily be linked to form complete journeys based on the difference between the subsequent trip's boarding time and the previous trip's alighting time for each bus passenger. By comparing the interchange time and the connecting bus route's headway, this research also provides a way to evaluate connecting bus services and bus passengers' interchange patterns.

Finally, this research can be expanded to the full bus network and other travel modes, opening the door to developing more comprehensive data bases for use in intermodal network planning.

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

Thesis Supervisor: John P. Attanucci
Title: Research Associate of Civil and Environmental Engineering

Acknowledgments

I would like to thank Prof. Nigel Wilson and Mr. John Attanucci for their guidance, support and attention to detail during my two-year graduate study at MIT. Through working with them, I benefit a lot not only on research skills and aptitudes but also on their industriousness and self-discipline. I like the weekly Wednesday afternoon meeting, when we three sat in Nigel's office discussing the research work. It is a time of wisdom and innovation, which motivates me to move further.

I am grateful to the staff at Transport for London (TfL), especially Rosa McShane, John Barry, Daly Aiden, Lisa Labrousse and Hélène Bataille for providing the great support for this research. The two-month summer intern in TfL provided me invaluable field experiences in the operation of public transit. Thank you!

My officemates, Mr. Frumin and Mr. Ehrlich as well as Jinhua provided indispensable technical assistance. My best friends, Wow Fred Webb, Rizard, Candy Gaga, Ehrlich Boy, Frudog Cookie Monster, Jerato, Tony Iron Man, Hulian, Alburg, David Uniman and BRT expert Harvey have brought me all the great happiness and made my life in the US & A and in 123 & 5 more diversified and colorful. Thank you, my dearest friends!

I also would like to thank Ginny for all the help during my two-year study here. I still save the special *One-day Trip Guidance from London Booklet for Winnie* from you, Ginny. That is the best booklet I have ever got!

Special thanks to MIT Writing Center teacher Ms. Elizabeth Fox for helping me with the writing and for mentoring me during the past three months. I am so lucky to have such a great friend and mentor like you!

Most importantly, I would like to thank my parents and my sister: their support, encouragement, patience and love make me who I am and help me go through all the challenges. My parents have supported me unconditionally since I went to college. My dear mother, who is a brave and warm-hearted woman, has set up a model for me to follow for the rest of my life. I love you Mommy! And I know "Impossible is Nothing!"

Thank you all. Any remaining errors or omissions are entirely my own.

Table of Contents

List of Tables

List of Figures

1 Introduction

The ever-increasing usage of the Automatic Data Collection Systems (ADCS) generates new transport data that can be used by transport service providers for a range of applications. These data were previously too expensive to obtain by traditional manual surveys. This research is undertaken to assess the feasibility and ease of using this ADCS archived data to provide useful information for decision-making by different departments within a transit agency.

Specially, this research explores the application of the ADCS archived data to public transport planning with a focus on bus passenger Origin-Destination (OD) inference at the bus-route level and on travel behavior, using London as an example. To infer the OD matrices for bus passenger trips, the characteristics of available ADCS archived data are examined. With the estimated OD matrices and the Automatic Vehicle Location (AVL) data, the alighting time for bus passengers can also be estimated. Then, bus journey stages can easily be linked to form complete journeys based on the difference between the next trip's boarding time and the previous trip's alighting time. Thus interchange patterns and other travel behaviors can also be studied. This research can also be expanded to the full bus network and other travel modes, opening the door to providing comprehensive database for more effective support of intermodal network planning. This chapter presents the research motivation, objectives, and approach, as well as an overview of London's bus network planning process.

1.1 Motivation

Public transportation systems provide people with communal transportation services, generally by bus or rail. Public transportation provides access to those who cannot or do not choose to drive, congestion reduction by providing a high-quality alternative to the automobile, and land use influence by allowing agglomeration of economic activity in cities. Finally, as an environmentally friendly alternative, public transportation has obtained great support from the government and the public.

There are high public expectations for the services that public transportation systems can provide. In order to provide better service needed to attract choice riders and generate a greater return for the public support as well as maintain success in the future, public transit agencies have gone through organizational changes, including greater operating staff responsibility, accountability and increased customer orientation. They have also implemented new technologies for better information provision and more effective real-time operations control as well as improved vehicle design.

For the provision of better services, service planning is one of the most important support functions. Major planning elements include data collection, problem identification, ridership estimation and other variables. For service design, most agencies have guidelines for scheduling based on maximum (policy) headway and maximum passenger crowding. When transit agencies set the standards, peak load at the maximum load point is particularly important for ensuring that adequate capacity is provided. Research shows that increased understanding of passenger behavior can be used to make gains in efficiency, customer service and cost reduction for the agencies involved. Among them, the passenger OD matrix provides a basis for evaluating the performance of individual bus routes and the bus network as a whole. Reliably estimated OD matrices can be used in service planning, in operations analysis, in before-and-after impact analysis, and in service management.

Traditionally, transit agencies obtain OD matrices by conducting occasional on-board passenger surveys and using various techniques to expand the survey results based on manual boarding and alighting counts at stops. However, such passenger surveys are expensive to conduct and therefore are extremely infrequent. The sample size will also be limited due to the high cost of data collection.

In the past decade, Automatic Fare Collection (AFC) systems have become increasingly popular as they provide an efficient and cost-saving alternative to traditional manual fare collection methods. Some research has pointed out and evaluated the potential benefits of using AFC data for transportation planning, especially using the AFC data to obtain OD matrices for performance measurement and service planning (Barry et al., 2002). Due to the limitation that most AFC systems record the passenger trip boarding location coarsely at the bus-route level, it is still difficult to get detailed information about where individual passengers board a bus. Integration of the AFC system archived data, which includes characteristics of each fare card transaction, with the AVL system data, which includes vehicle locations, offers a solution by matching the vehicle location information against the passenger trip information to help transit planners to infer the individual passenger's boarding bus stop (Zhao 2006).

Research has also been conducted on utilizing the ADCS archived data to infer the destination for an individual bus passenger's trip. Current system-wide or even single route level OD matrices are usually not available in most transit agencies. One of the primary advantages of using the ADCS archived data is to make these OD matrices available to transit agencies. Also, in terms of reduced cost, larger sample size, larger time span coverage and a more automated system, the ADCS provide a good way to obtain daily OD matrices (for both weekdays and weekends), which is very helpful for public transportation network planning, especially for large, complex and dynamic networks like London's.

Interchange planning has long been recognized as a key component of public transportation planning. It reduces operational costs and supports efficient route planning in exchange for passengers' inconvenience. Making interchanges less burdensome must be a critical consideration in public transportation planning. Acquiring interchange data is very difficult as it adds more possible variables to the data collection efforts. Before the advent of the AFC system, the only way to get such data was through a manual survey of passengers about their itineraries. Some research has been performed on bus passenger interchange behavior using AFC data to link journey stages into complete journeys based on elapsed time thresholds. Nonetheless, this kind of research work has usually been limited because the spatial scale of bus boarding locations was only at the bus route level and there was no information on the bus passengers' alighting times or locations. Therefore, the interchange time defined in such research is actually the bus journey time plus the "true" interchange time. However, with the implementation of the AVL system, the location and time for each bus can be obtained, which provides a new way to infer individual bus passenger's boarding and alighting locations as well as alighting times. Hence interchange times can be estimated more accurately by calculating the difference between the alighting time on the previous bus trip and the boarding time on the next transit trip.

1.2 Research Objectives

The objective of this research is to make the best use of the ADCS archived data to support public transportation service planning. This goal will be achieved by assessing the feasibility and ease of using these ADCS archived data to provide useful information for decision-making in service planning using London as an example. The following questions will be addressed:

1. What kind of passenger behavior information can be obtained from the ADCS to use in public transportation service planning?
2. How can this information be used for the decision-making within transit agencies?

Answering these questions will provide a better understanding of what ADCS can contribute to service planning by making more effective use of the available data. This research can also be expanded to apply to any transit agencies with similar ADCS.

1.3 Research Approach

In order to test the hypothesis that the ADCS archived data can be used to improve bus service planning, two data samples are used: (a) a 100% sample of AFC data for a sample of Transport for London (TfL) bus routes for two weeks, and (b) a 100% sample of AVL data for two weeks for all the routes that are parallel to or intersect the selected bus routes. The general approach is based on constructing trip chains for individual bus passengers. The methodology used here involves data cleaning and integration, inference of bus passenger trip boarding location, and inference of bus passenger trip alighting location and time, as well as interchange pattern analysis. The inference results are further validated by comparing with the BODS manual survey data.

Since this thesis applies this methodology to the bus network in London, the next section presents a brief overview of London Buses network and service planning.

1.4 Overview of London's Bus Network and Service Planning

This thesis uses the London bus network as a basis for analyzing the application of the ADCS archived data to bus network planning with a focus on OD estimation and travel behavior, especially the interchange pattern analysis. This section discusses the basic characteristics of the London bus network and service planning.

London has one of the largest bus networks in the world, with over 6 million passengers transported on its 700 routes every day. In a report by the Mayor of London (2009), it is reported that "Bus usage is growing at its fastest rate since 1946. More than two billion passenger trips were made on London's fleet of more than 8,000 buses in the year to March 2009. The number of operated kilometers has also risen to 478 million, the highest since 1957".

Every 5-7 years, the Bus passenger Origin and Destination Survey (BODS) is conducted by TfL for each bus route in London. This survey obtains detailed information about passenger travel patterns, including the number of people boarding and alighting at each stop, the purpose of travel, the location of the boarding and alighting of each journey, and how passengers get to the bus boarding stop and from the bus alighting stop to their final destination. Expansion factors are added to account for non-returned survey cards and non-surveyed bus trips. An automated database (BODS database) is compiled with the survey results. The automated reports from the BODS database include boardings, alightings, and loads at each stop (or stop zone) along a route, as well as its parallel routes. BODS can be considered the primary data system used by Network Development in TfL because it provides the detailed, disaggregate passenger demand information that is necessary for bus network planning in London. A major limitation of this type of survey is that it records passenger travel for only one day per route. Recognizing the substantial network growth and changes, supplementary data from other sources are needed for network planning. Moreover, although surveyed passengers are asked for their ultimate origin and destination in addition to their travel on the route itself, this information is rarely transferred from the paper surveys into the BODS database and is therefore not readily available to network planners.

In addition to BODS survey data, London bus planners also get timely route-level passenger demand information from Electronic Ticketing Machine (ETM) transactions, which are downloaded from bus garages to the Bus Contracts Management System (BCMS). One drawback of this data, however, is that it only records aggregate ridership for each bus trip, while detailed information for each bus passenger trip cannot be obtained directly.

The Oyster smartcard was launched in London by TfL in December 2003 as a new ticketing medium. It is now accepted on the Underground, buses, the Docklands Light Rail (DLR), Tramlink, and National Rail stations. Though this data source has not yet been fully used by London bus network planners, the Oyster Smart Card data is readily available. It has large sample sizes and offers a network perspective. Bagchi and White (2004) summarize the key benefits of Oyster Cards in terms of planning as:

(1) much larger volumes of individual passenger trip data than it is possible to get from manual surveys;

(2) ability to link individual passenger trips to individual cards or travelers;

(3) continuous trip data covering longer time periods than manual surveys;

(4) proportion of different types of customers using transit services. (2004)

In addition, using the Oyster smartcard data enables one to determine linked trips and the ultimate OD flows across the network. This process can be repeated on a day-by-day basis to assess variability in trips and get more accurate estimates of ridership for specific days of the week and times of the year. It provides an easier and more reliable way to get more detailed passenger behavior information than manual survey data, which can help transit agencies to make gains in efficiency and reduce cost.

1.5 Thesis Organization

This rest of thesis is organized into five chapters. Chapter 2 proposes the general transit passenger OD inference methodology as one way to obtain useful information from the ADCS archived data for public transportation planning. This chapter begins with a review of the previous research on inferring OD matrices using the ADCS archived data from transit networks. It further discusses the general OD inference methodology that can be applied to any transit agencies having similar ADC systems.

Chapter 3 presents the bus passenger OD inference results for the selected bus routes in London. It begins with a detailed description of the data sources that are used in this research. Then it moves to the inference results for the selected bus routes in

London. Some data issues related to terminal bus stops during the OD inference process are discussed, followed by a recommended resolution.

Chapter 4 focuses on the validation of the OD inference methodology that we apply to London buses. It starts with a primary analysis of BODS, ETM and Oyster smartcard data for all the surveyed bus trips, after which, origin and destination inference results are validated separately by comparison with BODS data for all the surveyed bus trips. This chapter concludes that the inferred OD matrices are generally valid and provide more detailed information for a wide range of operating conditions.

Further applications of this methodology and the inferred OD matrices are discussed in Chapter 5, which includes daily ridership, load profile and trip distance distribution for weekdays and weekends for different time periods and different directions for the selected bus routes. A further study of interchange times and the subsequent travel behaviors of bus passengers are also presented.

Chapter 6 concludes with a summary of the key findings, discussion of possible improvements to enhance current practices, and suggestions for future research.

2 Origin-Destination Estimation Methodology

This chapter describes the proposed Origin-Destination estimation methodology in detail. It begins with a review of the previous research on using Automated Data Collection Systems (ADCS) to create Origin-Destination (OD) matrices for public transportation networks. Then it discusses the characteristics of ADCS archived data required in order to infer transit passengers' origins and destinations, followed by a description of the general OD estimation methodology that can be applied to any transit agencies with similar ADCS.

2.1 Previous Research on OD Estimation using ADCS Archived Data

Cui (2006) has summarized OD estimation techniques using manually collected data. Basically, the OD matrix can be obtained either by surveys or by techniques which combine various sources of data. The increasing use of ADCS provides a cost-effective way to collect data that goes far beyond the scope of data collected from manual surveys. Although most ADCS are designed to support specific narrow functions, the resulting data can be applied to areas far beyond their design purposes. This thesis illustrates the potential for the ADCS to provide transit agencies with new, richer data sources at low marginal cost. For the purposes of this thesis, only the most important attributes of ADCS archived data are described in detail; for example, the recorded location and time of each farecard transaction, which offer a new way to infer passenger trip ODs.

Although ADCS change the quantity, type and quality of data available to transportation planners, there is still a gap between what these systems can directly provide and what transit agencies need in practice. One example is the passenger trip OD matrix, one of the most important elements for transit planning, which cannot be obtained directly from the ADCS since most AFC systems record entries but not exits. Some research has been carried out to bridge the gap between what ADCS archived data offers directly and what is desired by transit agencies through extensive data analysis using database and GIS techniques.

Recent passenger trip OD estimation research can be divided into three cases: both entry and exit locations are recorded by the ADCS; only entry locations are recorded; and neither entry nor exit locations are recorded. The following sections review approaches to deal with these three cases.

2.1.1 OD Estimation with Both Entry and Exit Locations Recorded

An early example of an automated fare collection system was the Bay Area Rapid Transit (BART) system. Buneman (1984) initiated the creation of OD matrices by using farecard-based data in the BART system. In that case, the distance-based fare structure required identification of both entry to and exit from the system, which provided a more accurate passenger trip OD matrix with simpler processing than for entry-only systems.

Jang (2010) further examines the possibilities of using data from such closed systems for transportation planning applications in the city of Seoul. One feature that distinguishes the Seoul ADCS from many other cities is that it records each trip's entry and exit times and locations, as well as the trip chains with interchanges. Based on this dataset, Jang analyzes interchange patterns and identifies interchange points that need improvements by examining the points where interchange demand exceeds 5,000 per day and/or the average interchange time exceeds 10 minutes.

2.1.2 OD Estimation with Entry-Only Boarding Locations Recorded

Trépanier et al. (2007) present a model to estimate the alighting stop for each individual passenger boarding a bus with a smart card in Gatineau, Quebec. In their case, the ADCS store the location where the passengers boarded using the onboard positioning equipment. Hence, they only need to develop a way to infer the destination in order to define the passenger trip. The destination inference is based on two assumptions: first, the destination of a passenger's trip is the first stop of his/her following transit trip ("normal trips") and second, the passenger's last transit trip returns to the stop where he/she first boarded ("last trips"). They report a success rate of 66 percent in the first application of this methodology but up to 80 percent in peak periods. Figure 2-1 illustrates the destination inference process.

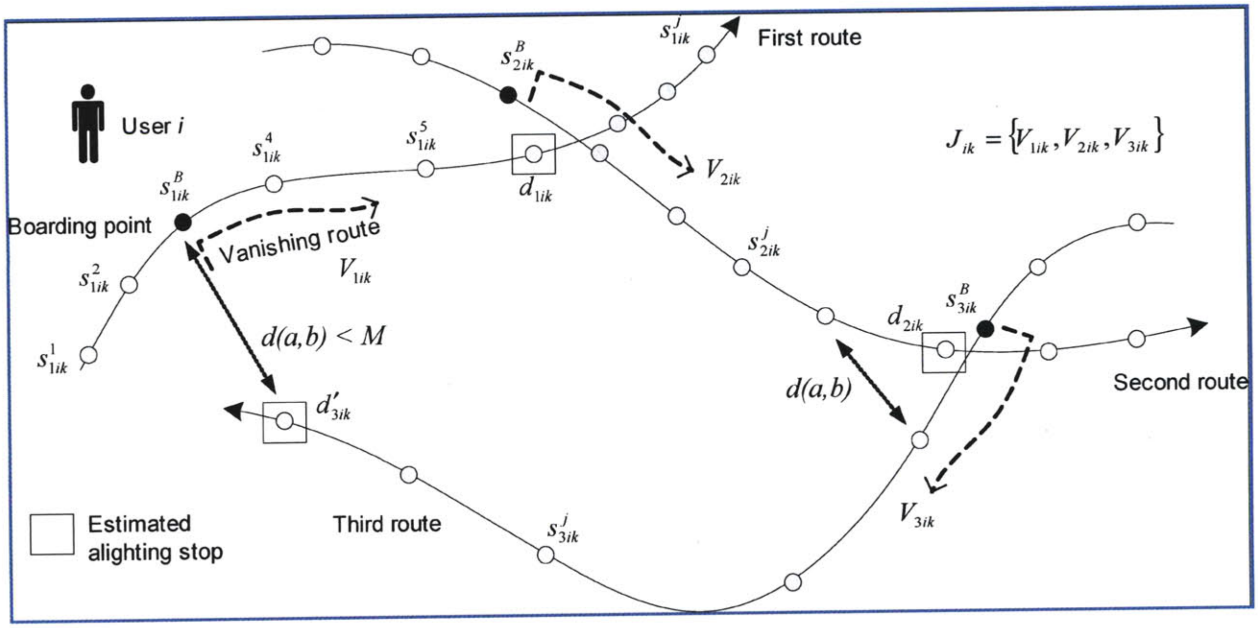


Figure 2-1: The Alighting Location Estimation Model for "Normal" and "Last" Trips (Trépanier 2007)

2.1.3 OD Estimation with Neither Entry Nor Exit Locations Recorded

If neither the boarding nor the alighting locations can be obtained directly from the ADCS, then both of them need to be inferred. Methodologies have been developed to infer the boarding and alighting locations separately using ADCS archived data as illustrated below.

a) Origin Inference

Depending on the data available, boarding locations can be inferred in two ways. One is to infer the origin with an AVL system, and the other is to infer the origin without an AVL system. Clearly, the origin inference process with an AVL system is more reliable.

- Origin Inference with an AVL system

In the Chicago Transit Authority (CTA) case, the AFC system is an entry-only system, meaning passengers only swipe their cards when entering a rail station or boarding a bus, so no information about the destination is provided directly. In addition, the boarding location is only coarsely recorded at the bus route level with no specific information about the bus stop provided. Zhao (2006) proposes a way to integrate the AFC and the AVL systems to infer boarding locations. He integrates the two systems by matching the vehicle location information from the AVL against the passenger trip information from the AFC to infer the individual passenger's boarding stop. The AFC transaction time for a passenger boarding should be shortly after the corresponding AVL time recorded as illustrated in Figure 2-2:

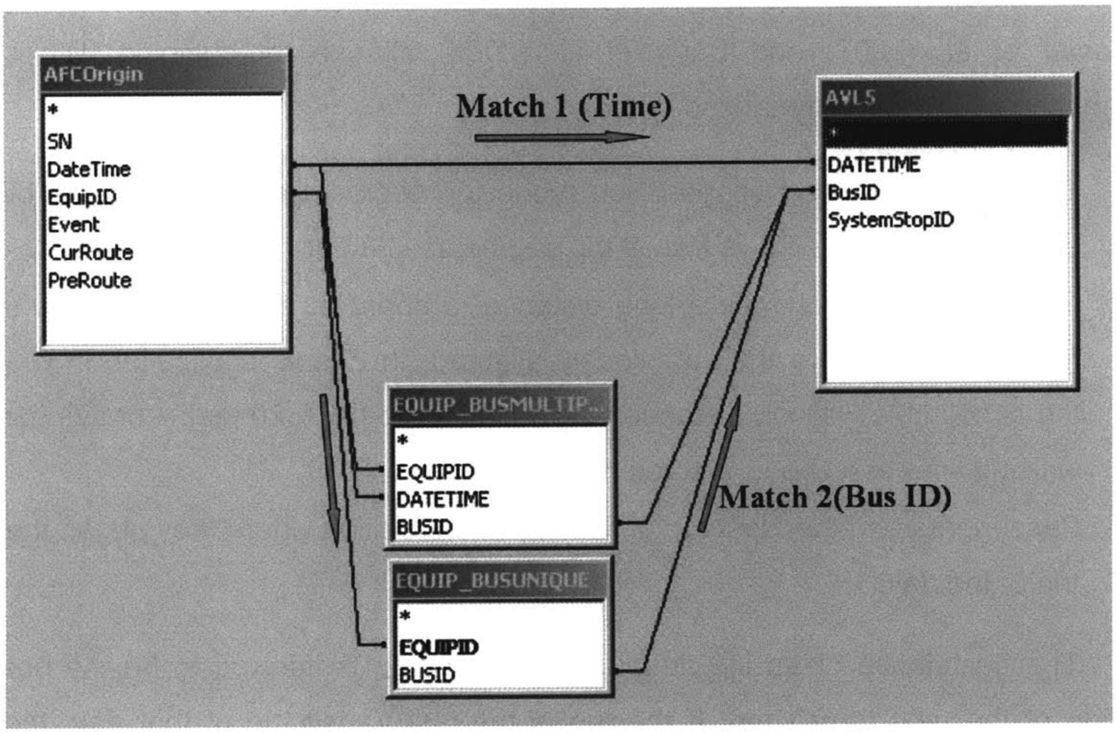


Figure 2-2: Combining AFC with AVL to infer the boarding stop (Zhao 2006)

Cui (2006) further applies this origin inference methodology to the bus network in Chicago, beginning at the single route level, and then at the network level with an origin inference rate at 90% for all CTA passenger trips using AFC cards.

- Origin Inference Without an AVL system

Barry et al. (2008) propose an approach to identify the specific boarding and alighting locations using only the AFC data with no location information in New York City. Without an AVL system, it is impossible to know the accurate bus location at any time. They use the scheduled run time to estimate the location of a bus along its route at the time of the AFC transaction. Transfer information obtained from the AFC system is used to adjust for buses running off schedule. For example, if a passenger has two bus trips successively within a short period of time and the two bus routes intersect, then the second bus must be close to the intersection stop at the transaction time, and the scheduled bus trajectory can be adjusted accordingly.

b) Destination Inference

To infer the destination for individual passengers' trips, Zhao (2006), Cui (2006), Trepanier et al. (2007) and Barry et al. (2008) all use trip-chaining with similar assumptions as those summarized by Zhao (2006):

- There is no private transportation mode trip segment (car, motorcycle, bicycle, etc) between consecutive transit trip segments in a daily trip sequence;
- Passengers will not walk a long distance to board at a different rail/bus station from the one where they previously alighted. In Zhao's application in Chicago, the acceptable walking distance was assumed to be 1320 feet -- or five minutes' walking time at a speed of three miles per hour;
- Passengers end their last trip of the day at the station where they began their first trip of the day.

The destination of an individual passenger's trip is inferred to be the boarding location of the next transit trip. If the trip is the last transit trip of that day, then the destination is inferred to be the boarding location of the first transit trip of that day. Barry et al. (2008) further assume that travelers making only a single trip on a day have the same destination distribution as multiple-trip card users, given the same boarding location. They validated their results by comparing them to a travel diary survey for subway riders. This study found that this methodology resulted in 90% valid destinations.

Cui (2006) further applies this destination inference methodology to the bus network in Chicago with an inference rate of 67% for all the transit passenger trips using AFC cards. The network OD matrix is formed as shown in Figure 2-3:

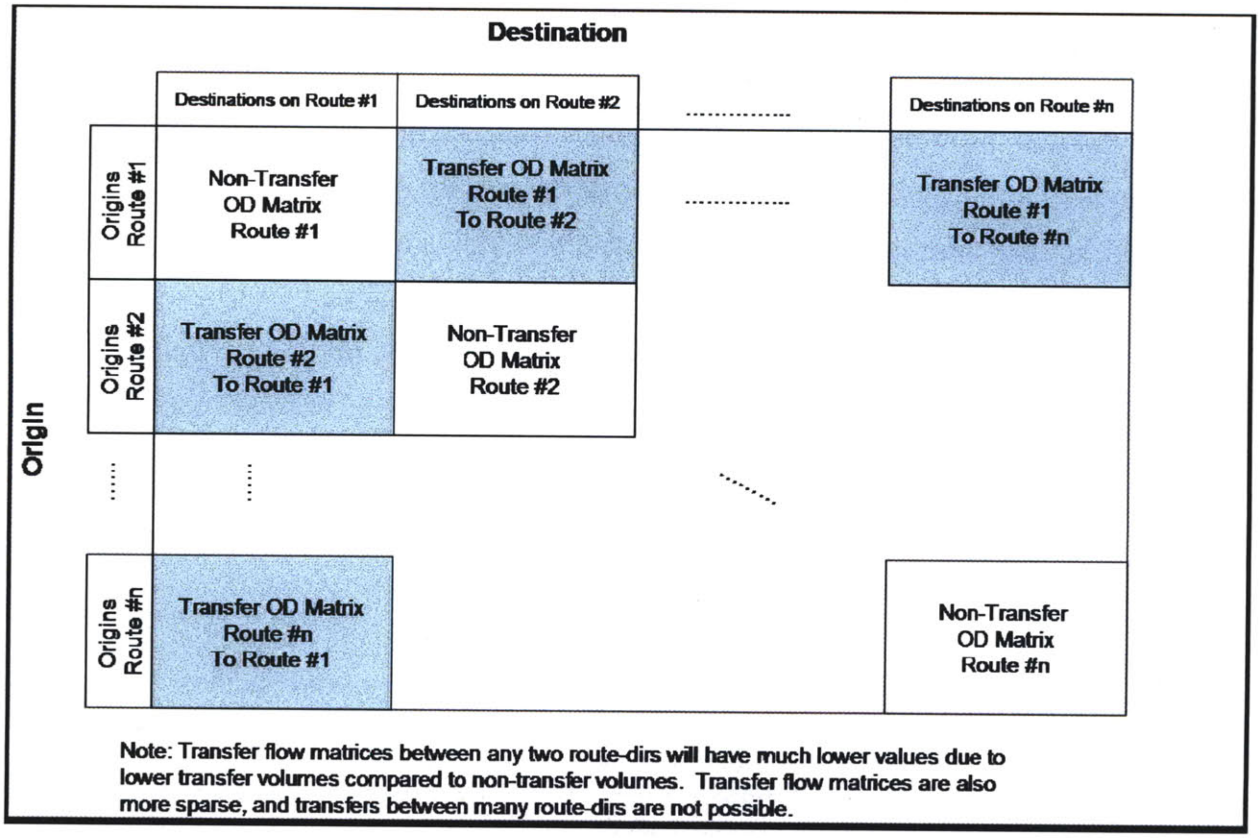


Figure 2-3: Network level OD matrix (Cui 2006)

2.2 General Methodology to Infer Transit Passenger Origins and Destinations

The transit passenger OD estimation methodology described here is built upon the trip-chaining OD estimation method applied in Chicago in 2006 by Cui (2006). Since different transit agencies may have different data sources with different characteristics, this section describes a general methodology that can be applied by any transit agency with ADCS, regardless of the specific characteristics. It begins with a discussion of the required characteristics of the ADCS archived data, followed by a detailed description of the methodology used to infer transit passengers' origins and destinations.

2.2.1 ADCS Data Preparation

ADCS are among the most powerful ITS technologies available to transit planners. However, many ADCS were not initially designed for data collection purposes. For example, ADCS are often designed to announce bus stops, to collect fares, or to report emergencies. This often poses problems in using the data, especially because much of the data collected may be intermittent and fragmented. These limitations make the raw data meaningless without extensive analysis. In addition, different systems supplied by different vendors often have incompatible data structures and the data are often stored and managed in different database management systems. All of these factors make the integration across different systems difficult and severely limit easy access to the ADCS archived data.

Though different transit agencies have different ADCS, in order to implement the proposed OD inference methodology, some basic characteristics of the ADCS archived data are required. In general, three data sources are necessary: AFC, AVL and GIS data.

a) AFC Data

An AFC transaction record is generated each time a passenger swipes (or taps) a farecard at AFC equipment. The record includes the time, location (bus route and trip), and a unique serial number for that farecard. The time recorded is when a passenger swipes (or taps) a farecard to board a bus or enter a rail station to wait for the train; the location recorded may be as coarse as the bus route ID and the bus trip ID, in which case the actual boarding location cannot be obtained from the record directly; the unique serial number is assigned to each farecard to track each card holder's multiple trips over the life of the farecard. Since most of AFC systems are entry-only, passengers only use their farecards when boarding a bus, and no information about passengers' destinations can be obtained directly. The basic information needed from the AFC system for the OD inference includes:

- time (date and time for the farecard transactions);
- bus route ID, showing the route number that each farecard holder boards; and
- bus trip ID, showing the trip number of the bus, which also indicates the direction for each trip;
- farecard ID, unique to identifier for each farecard holder.

b) AVL Data

The AVL system records vehicle location information, which may be used to keep the driver informed of on time performance, to inform the bus control center of current bus location, and to drive the automatic stop announcement system.

An AVL record typically includes the following data:

- the vehicle's progress along a route, for instance, "the arrival" and "the departure," including a time-stamp (always to the second);
- the location information at each bus stop; and
- the identification information such as the bus route number and the bus trip number, which are the most important features needed to infer bus passengers' boarding locations.

c) GIS Data

To infer the alighting locations, we also need GIS files encompassing the bus and railway networks, as well as the road network of the studied area. These GIS files should include:

- every bus route in the network;
- each direction for every bus route, as routing varies by direction;
- every bus stop/rail station in the network.

Although all of these data are collected automatically, possible human errors could still be involved in the AFC and AVL systems. For example, bus drivers may not log on/off the AVL system correctly. And bus trip number and direction may not be changed at the right time, which may cause problems at the terminal stops. Using the published schedule, it is possible to identify the correct trip that the given transaction should be assigned to. These issues will be discussed in detail in Section 3.2.

2.2.2 Origins Inference

The basic idea for inferring bus passengers' origins is that it should be possible to identify the boarding stop for every passenger with a farecard who boarded a bus with a functioning AVL system. For a given route and trip, the fare collection timestamp from the farecard is used to search through the AVL data to determine the boarding stop.

This step is logical and straightforward and can easily be implemented by matching the farecard transactions against the AVL recorded bus run information to identify bus stops where passengers board.

2.2.3 Destination Inference

The destination inference method is built on the establishment of trip-chains for each farecard user, with the same assumptions described in the above literature review. Before implementing the methodology, lookup tables are needed to help infer the alighting stop which is nearest to the next trip's boarding location. Lookup tables can be derived by calculating the distance between bus stops and between bus stops and rail stations with up-to-date GIS files.

The detailed procedures to implement this methodology are illustrated in Figure 2-4. Assume the bus trip segment currently under examination is trip segment ***s*** by person (farecard) ***p*** on day ***d***, and its subsequent trip segment is ***ss***. The algorithm determines whether trip segment ***ss*** is on bus or on rail and moves onto the "next trip" rule with a lookup table for bus sub-procedure or "next trip" rule with a lookup table for rail sub-procedure respectively. There are two exceptions as also indicated in Figure 2-4:1) if trip segment ***s*** is a single trip made on a day, then the destination cannot be inferred; 2) when trip segment ***s*** is the last trip segment of the day, the first trip segment of the day is regarded as the trip segment immediately following ***s*** and then the "next trip" rule can be applied to infer the destination of the trip segment ***s***.

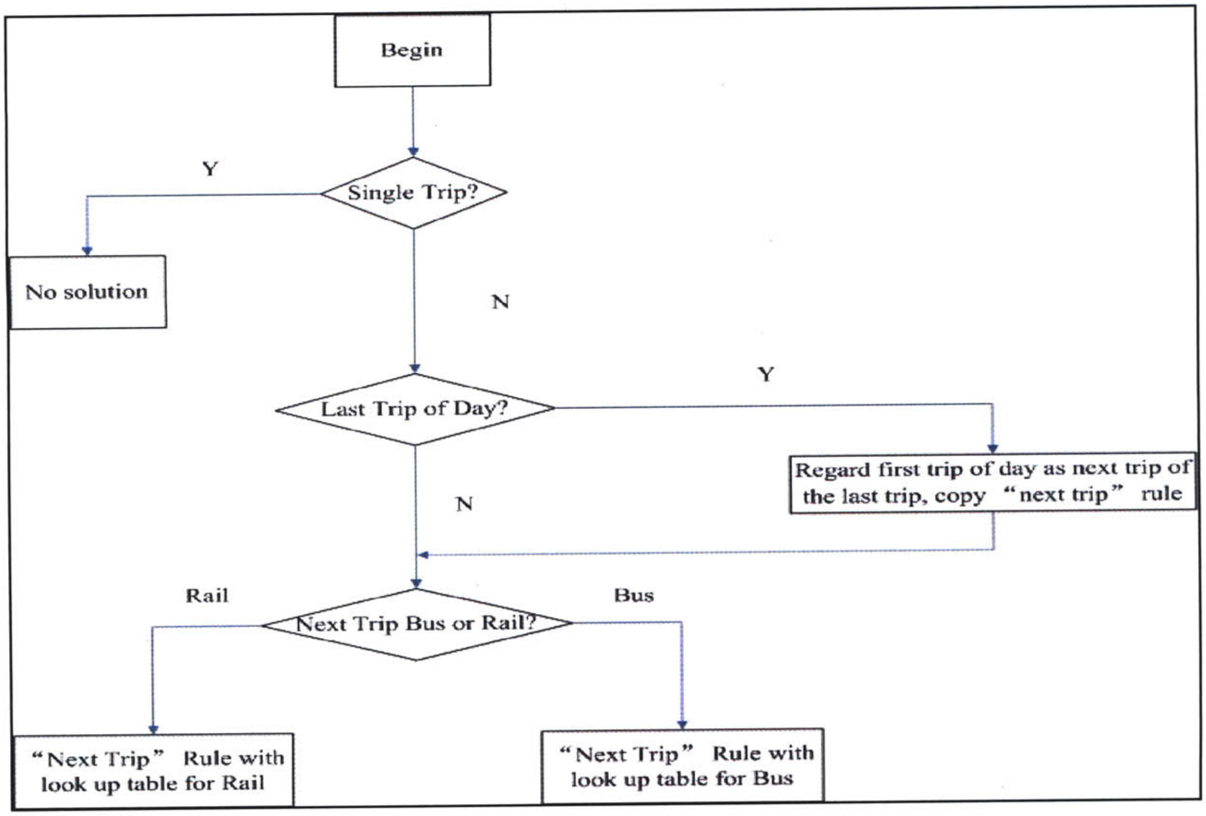


Figure 2-4: Process for destination inference

2.3 Summary

From both the literature review discussed in Section 2.1 and the methodology described in Section 2.2, it is feasible to infer the transit passengers' OD matrix. Generally, for transit agencies equipped with ADC systems, as long as they have the AFC, AVL and GIS data with required characteristics as described in Section 2.2.1, the methodology proposed in Sections 2.2.2 and 2.2.3 can be applied to obtain the passengers' OD matrices automatically.

Moreover, since the ADCS have extensive spatial coverage and a full temporal span of continuous data, detailed daily transit passenger travel information can be obtained easily. With traditional methods, only key locations and "typical" day and peak hours are studied, but with ADCS, transportation planners can get much more useful and reliable information about the transport system with continuous data. They can also get a complete picture of the travel demand variations for different days, different time periods and different locations, which can help them to make better service planning decisions.

3 London Case Study: Origin-Destination Inference

Following the general methodology described in Chapter 2, this chapter develops a method specifically for London buses using data available at Transport for London. The characteristics of ADCS archived data in London and the feasibility of using these data to infer bus passenger OD matrices are discussed in detail in Section 3.1. Five bus routes in London are selected for study with results presented in Section 3.2 and 3.3. The bus terminal issues encountered during the process are discussed in Section 3.2 as part of the discussion of origin inference. This chapter ends with a summary of the application of this OD estimation method to London.

3.1 TfL ADCS

There are four key data sources available in London for inferring bus passengers' OD matrices: Oyster Smart Cards, iBus, Electronic Ticketing Machines (ETM) and GIS with characteristics as described below.

3.1.1 Understanding Data Quality and Characteristics

(1) Oyster Smart Card Data

The Oyster Smart Card was launched in London by TfL in December 2003 as a new contactless ticketing medium. The penetration of smart cards is crucial for automatic OD inference as practiced in this research. It is currently estimated that about 80% of all journeys are made each day on the bus, Tube, DLR and London Overground using Oyster Cards (Transport for London, 2010). Oyster Smart Cards in London are owned by individuals and record every transaction the card holder makes in the public transportation system. For the Underground and Overground networks, both the time and rail station for entry and exit are recorded, while for bus, only the time of bus boarding and route information is recorded. Several types of analyses are possible with the smart card data, including ridership monitoring, revenue estimation and service performance measurement in the rail network. The key contribution of this research, however, is to develop a methodology to infer the origins and destinations for bus passengers in London using the Oyster smart card data and to develop related applications for the London bus network. For the purposes of this project, the following data are used:

Prestige (Oyster Card) data from November 8, 2008 to November 21, 2008, and from May 11 to June 7, 2009 for the entire TfL network including the following data fields:

- Daykey – identifies the date of the Oyster transaction
- Pid_encrypt – unique card ID, encrypted to provide anonymity and protect privacy
- Sequenceno – sequence number of the transaction
- Subsystemid – distinguishes bus trips from rail trips in the database
- T_cen[1] – transaction time (in minutes after midnight)
- Routeid – route number of bus transactions
- Bus_Trip_No – trip number of bus transactions
- Direction – bus route direction at boarding
- Next_Sequenceno – sequence number of the next transaction for the same Oyster card holder

Currently about 90% of bus passengers use Oyster Cards although this percentage is increasing over time.

To infer the alighting locations for bus passengers, we need the boarding information for bus passengers' next transit trips, which could be either by National Rail/Underground or bus. Before January 2, 2010, Oyster Cards could be used on National Rail only on trains travelling in Zones 1-9 (Transport for London, 2010). The coverage of National Rail network is increasing over time meaning that more information will be available in the future for destination inference.

(2) iBus Data

iBus is a £117m AVL and radio system which aims to help London Bus Services Limited (LBSL) to run more reliable and consistent bus service. The first installations took place in March 2007, with completion in April 2009. iBus data contain information about the

[1] The Oyster transaction times are truncated to the minute while the iBus AVL timestamps are to the second, and this will cause some problems when matching the two against each other to infer the origin. A method is proposed in Section 3.2.1 to deal with this issue.

route number and trip number as well as the direction for each bus trip, but most importantly, it provides a unique identifier for each bus stop and the observed departure time from each stop. iBus data for the selected bus routes and all parallel and intersecting routes for November 8-21, 2008, and May 11 to June 7, 2009 are used in this research. These datasets, provided by the Technical Support Group (TSG) of London Buses, contain the following fields:

- Shortdesc – service bus route number
- Direction – direction of trip
- Tripnr – LBSL trip number run by the vehicle
- Stoppointid – unique iBus identifier of each bus stop
- Stopsequence – sequence number of stops within the route at each direction, for example, 1 means the first stop at one direction along a bus route, and 2 means the second stop, etc.
- Shortdesc2 – LBSL bus stop code, which can be joined with GIS files
- Longdesc – LBSL bus stop name
- Observed Run Nr – run number of the observed trip
- Scheduled distance – distance between one bus stop and its previous stop in meters
- Sched Dist In Trip – cumulative distance of the trip from the origin stop to the stop in question in meters
- Scheduled departure time – scheduled departure time from this stop, including the date and time
- Observed departure time – observed departure time from this stop, including the date and time, but there is ambiguity about what is exactly recorded
- Tripid – unique iBus identifier of the trip

(3) Electronic Ticketing Machine (ETM) Data

Route-level passenger demand data are gathered through Electronic Ticketing Machine (ETM) transactions, which are downloaded from bus garages to the Bus Contracts Management System (BCMS). The BCMS is used to determine the number of journey stages over an entire day (24 hours) for any timeband (e.g., AM Peak usage from 7 to 9:30 a.m.) as recorded by bus drivers.

The ETM totals (the aggregate number of trips taken in each bus) from November 8 to 21, 2008, and from May 11 to June 7, 2009 are used in this research as control totals to expand the Oyster counts to all passengers.

(4) GIS Data

Up-to-date GIS files encompassing the TfL bus and rail networks are provided by Network Development of London Buses, which include the following files:

- Busnet_Routes071108 – a line file of every route in the bus network; for every route, a separate line indicates each direction, as routing varies by direction.
- Busnet_Stops071108 – a point file of every iBus AVL stop in the network.
- The railway network files and the London road network.

3.1.2 Methodology based on Oyster and iBus Data

Based on the characteristics of the Oyster and iBus data, the methodology proposed in Chapter 2 is applied. The basic idea is that it should be possible to determine the boarding stop for every passenger using an Oyster Card to board an iBus-equipped bus. For a given route and trip, the fare collection timestamp (including the date) from the Oyster smart card is used to search through the iBus data to determine the boarding stop. The next trip taken by the user is used to infer the alighting stop, where possible.

Refinements to the general methodology proposed in Chapter 2 are made due to the specific data characteristics in London. A feature that distinguishes the London case from other transit agencies is that in addition to obtaining the "seed" OD matrix from the Oyster and iBus data, London has Electronic Ticket Machine (ETM) data, which counts both the Oyster and non-Oyster passengers for each bus trip over the day. The ETM data is used to expand the "seed" matrix from the Oyster passengers to all passengers. Section 5.5.2 below discusses this issue in detail.

3.2 Origin Inference

Since the Oyster transactions only record the timestamp when every Oyster passenger gets onboard, and no information about the location is recorded, while the iBus AVL system records the timestamp when the bus door opens/closes at each bus stop for each bus run, the boarding stop can be determined by matching the Oyster transaction times against the iBus time data.

The origin inference procedure is implemented through a custom-built Java program, according to the assumptions, rules and limitations set out below.

3.2.1 Underlying Assumptions

Several assumptions are made to infer the origins:

- Since the Oyster transaction times are truncated to the minute, while the iBus timestamps are recorded to the second, the actual boarding times can be thought of as a random variable uniformly distributed over the 60 seconds starting with the recorded minute. When we match the Oyster transactions against the iBus timestamps, we first try to search for the iBus timestamp that is within the same minute as the Oyster transaction. If there is such an iBus record, then it is a perfect match for the Oyster transaction; if there is no such iBus record, then we set the expected value of the Oyster transaction time to be equal to the value recorded in the Prestige Oyster database plus 30 seconds and compare the time differences between the Oyster transaction and the iBus records.
- Each Oyster transaction time falls between two iBus transaction times, a "previous stop" and a "next stop." Fares are generally paid before the doors of the bus close. However, since the definition of "departure time" in the iBus system is ambiguous, one cannot say with certainty that an Oyster transaction should always be assigned to the "next stop." Experiments are made by matching the Oyster transactions to the "previous stop" and "next stop" separately and the effectiveness is compared as shown in Figures 3-1 and 3-2.

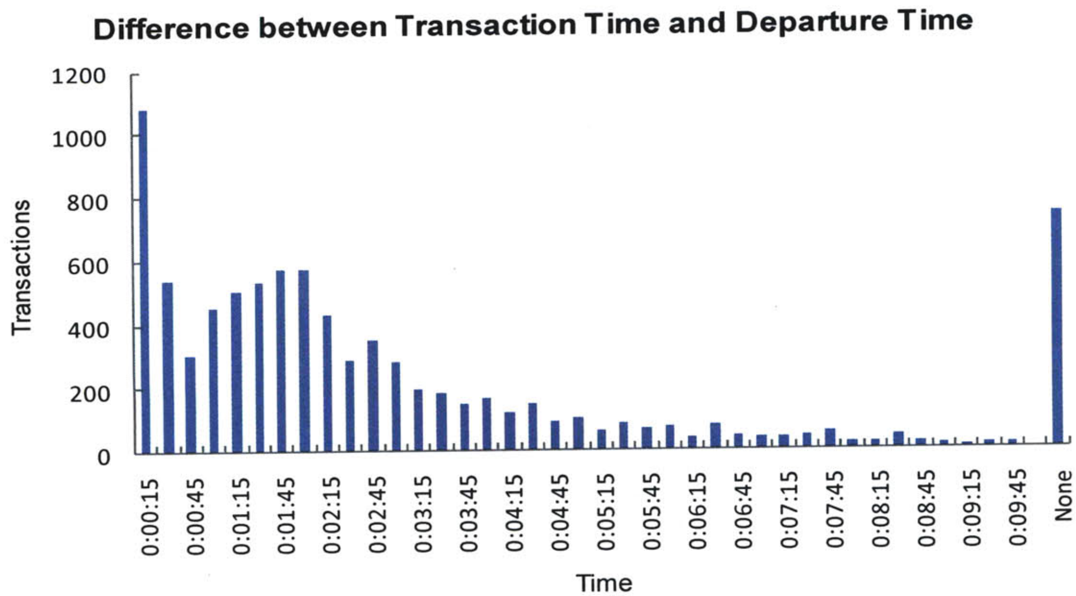


Figure 3-1: The effectiveness of "previous stop" rule (W4)

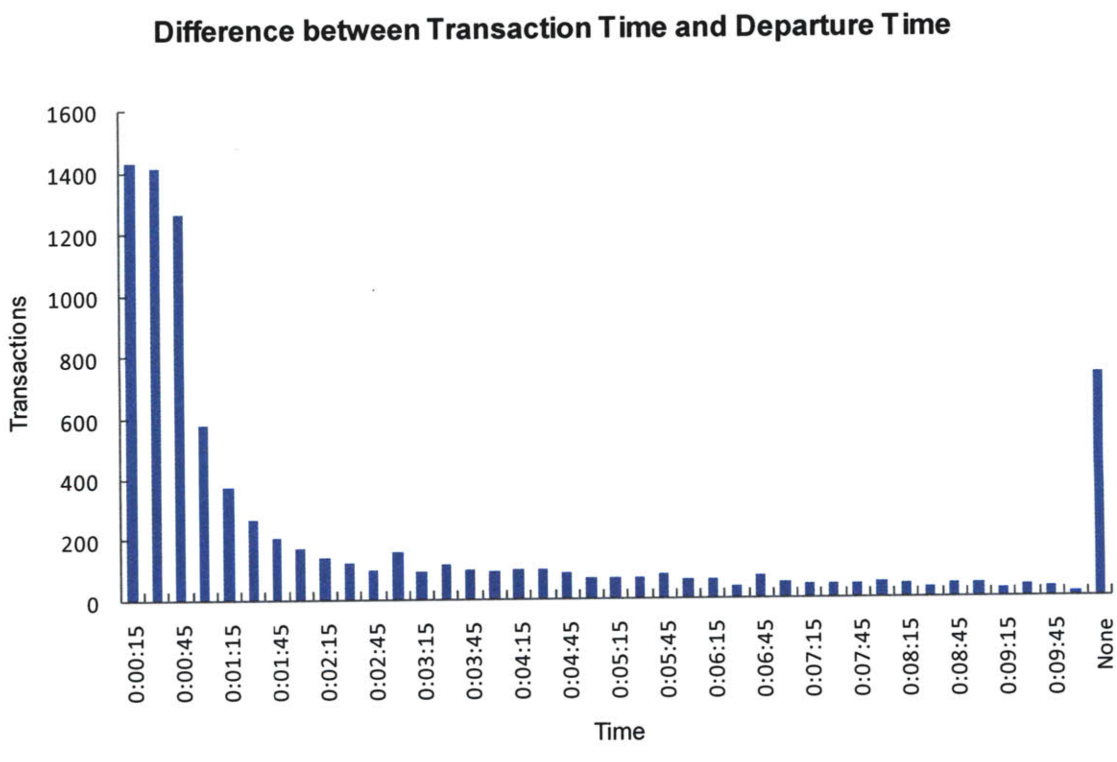


Figure 3-2: The effectiveness of "next stop" rule (W4)

Since both of these rules show some large differences between the Oyster transaction time and the iBus observed departure time, another rule, "closest stop" is also tested, as illustrated in Figure 3-3 below. "Closest stop," is the minimum of {(next stop – Oyster transaction), (Oyster transaction – previous stop)}, i.e., the closest stop in time to the Oyster transaction.

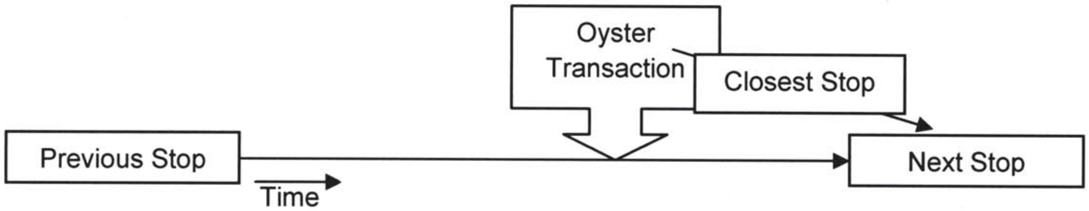


Figure 3-3: "Closest Stop" rule for origin inference

Figure 3-4 shows the effectiveness of the "closest stop" rule using Route W4 as an example:

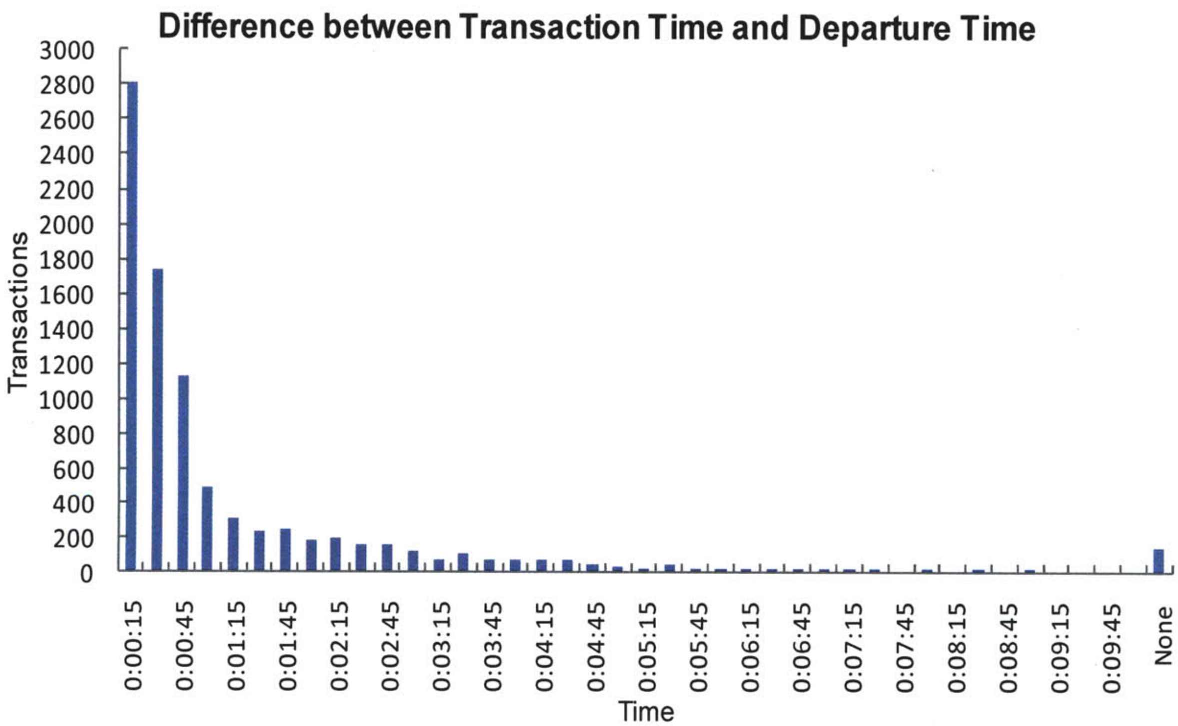


Figure 3-4: The effectiveness of "closest stop" rule (W4)

By comparing the three rules, the "closest stop" rule works best in minimizing the time difference between the Oyster transaction time and iBus observed departure time. Therefore, the "closest stop" is used to infer the origin.

In some cases, a bus driver may open the bus doors to allow boarding while waiting at the terminal for an extended period of time. In these cases, some passengers will board the bus a few minutes after the bus doors open. However, if the time between the bus departure and passenger payment is longer than a few minutes, it is likely that this data is due to an error and we cannot rely on it to infer the boarding stop. In this study, 5 minutes is used as the threshold to determine errors. If there is no valid iBus AVL record within this threshold value, this method considers the boarding stop for this Oyster transaction to be unidentifiable and this transaction record is discarded.

3.2.2 Process of Origin Inference

The Java program for origin inference operates as follows:

- Read all the iBus data, placing the data into the following hierarchical format (where "point" is a single record of a "departure" from a specific bus stop, including the departure time and the ID of the stop that the bus departs from):

  Route → Day → Trip → Point

  In this case, the "Route" level will contain the route ID selected for O-D inference and all of its parallel and intersecting routes for which iBus data are available. "Points" are ordered by iBus departure time, to make the search algorithm in the following step more efficient.
- Each Oyster transaction is then read. If a match is found for the route, the program then looks for a match for the date and trip. If a match is found, the transaction is matched to a stop (point) using the "closest stop" rule. If no match is found, that particular Oyster transaction is discarded.
- This process is repeated for all Oyster transactions in the file.

3.2.3 Origin Inference Results

Routes W4 and 70 are used here as examples to show the results of origin inference process.

- Route W4

Route W4 runs between Tottenham Hale (Ferry Lane Estate) and Haringey via Bruce Grove, Broadwater Farm, Turnpike Lane and Wood Green. This route is 7 miles (11km) long with daytime headways of 10 minutes and 20 minutes during the early morning and late night. This route connects with two National Rail stations and three Underground stations. The route schematic is shown in Figure 3-5 (the grey line).

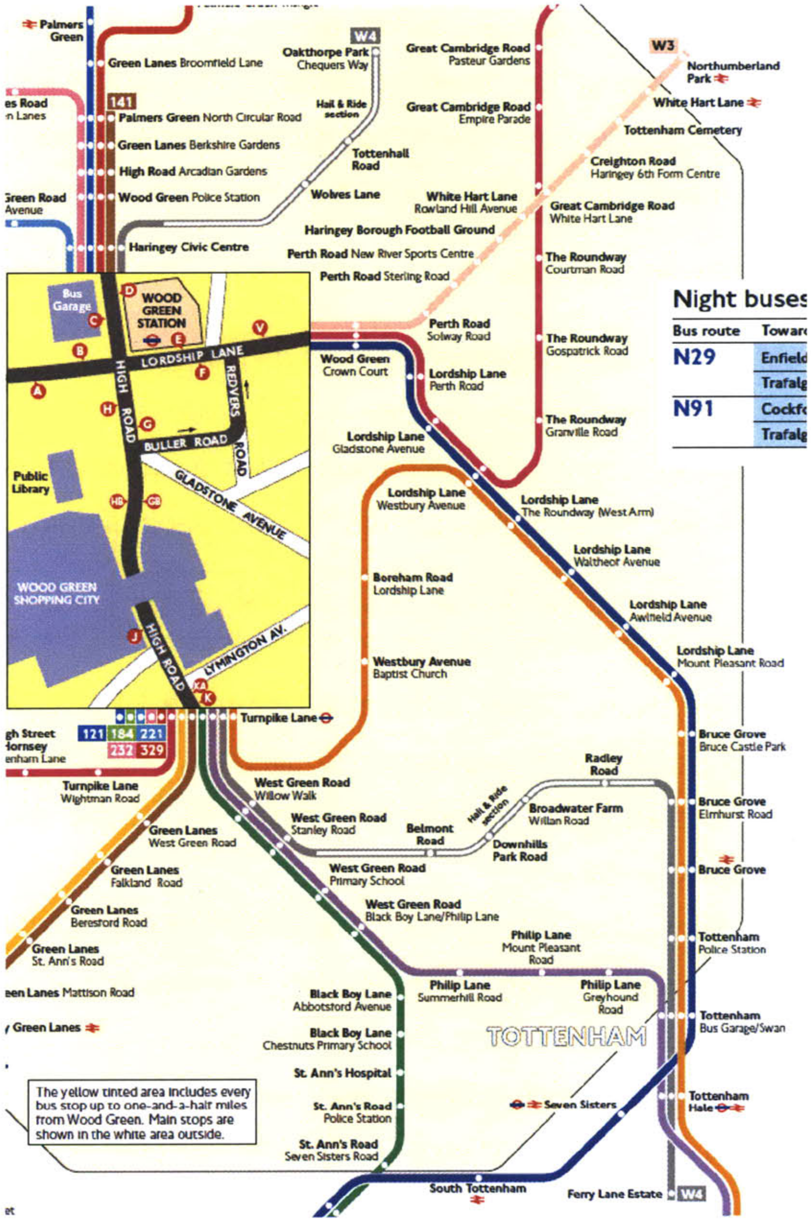


Figure 3-5: Route schematic for W4

(*http://www.tfl.gov.uk/tfl/gettingaround/maps/buses/pdf/woodgreen-2303.pdf*)

The W4 is also one of several bus routes in London which contain multiple "hail-and-ride" sections. These lengthy sections do not have any marked stops, and therefore no stops along these sections are recorded in the iBus system, and so they present a challenge when using the proposed methodology to determine the boarding and alighting locations for each passenger at the stop-level. In this study, the stops “bookending” the “hail-and-ride” sections are aggregated into a single “dummy” stop as shown in Figure 3-6. This way, passengers matched to a stop using the “closest stop” rule are assigned to a general “hail-and-ride” stop, not to a specific street intersection.

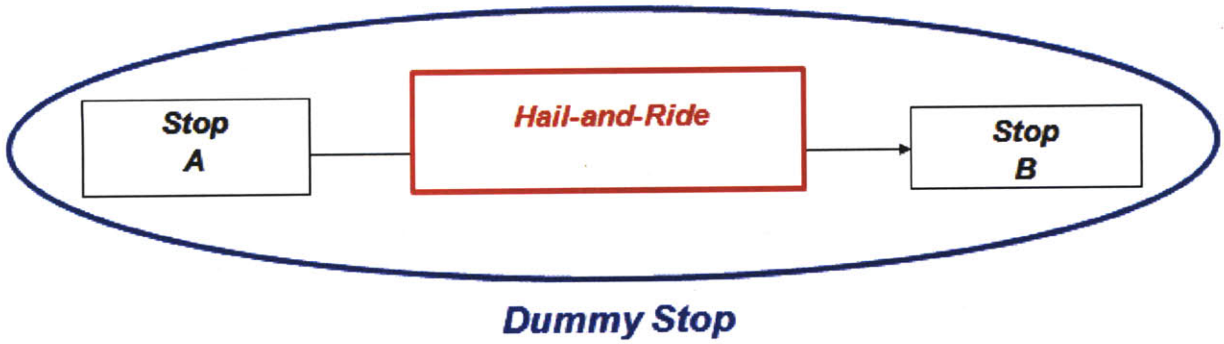


Figure 3-6: Dummy stop for “Hail-and-Ride” segments on Route W4

Of a total of 8585 Oyster transactions that are recorded on Route W4 on November 20, 2008, boarding stops were inferred for 94% or 8028 passenger trips. Most unmatched transactions are due to a lack of iBus data for the trip, since 6 out of the 192 bus trips run that day had no iBus data.

- Route 70

Route 70 runs between Acton and South Kensington via East Acton, Ladbroke Grove, Queensway and Notting Hill Gate. This route is 8 miles (12 km) long with daytime headways of 10 minutes and 15 minutes during the early morning and late night. There are no “hail-and-ride” segments for Route 70, which makes the matching less complicated. The route map is shown in Figure 3-7 (the pink line).

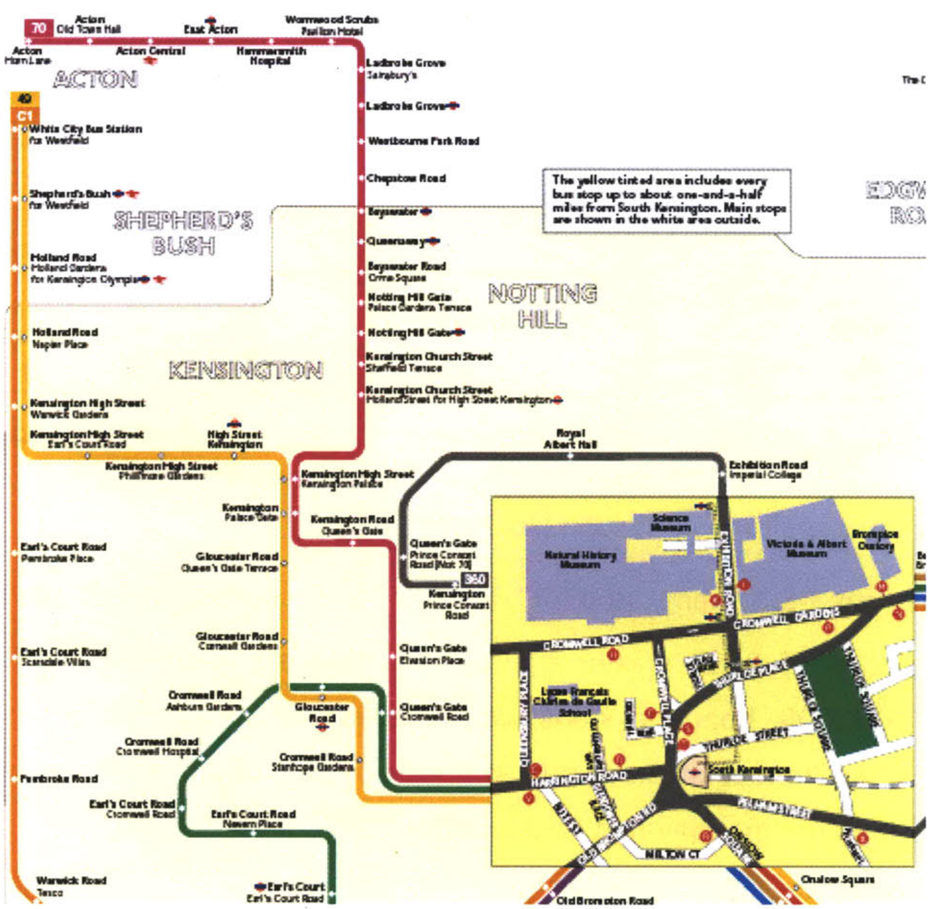


Figure 3-7: Route schematic for 70

(http://www.tfl.gov.uk/tfl/gettingaround/maps/buses/pdf/southkensington-2236.pdf)

There are a total of 196 bus trips on Route 70 with 8 bus trips lacking iBus data on November 11, 2008 (BODS survey day). A problem with the iBus system on this day is that there is no iBus data available for the first 5 stops westbound from South Kensington. Section 3.2.4 proposes a method to deal with this problem.

Of a total of 12074 Oyster transactions that are recorded on Route 70 on November 11, 2008, 11381 or 94% of the Oyster transactions have origins that were successfully inferred. Similar analyses are conducted for the other three selected routes. Table 3-4 at the end of Section 3.3 summaries the origin inference results for all five selected routes. In general, most of the routes have similar origin inference rates, with around 90% of all Oyster transactions having the origins inferred.

The major reason why some transactions do not have origins identified is a lack of iBus information for some trips. Another possible reason is that sometimes the timestamps recorded in the iBus system are inconsistent, which makes the Oyster transactions impossible to match against the iBus data.

3.2.4 Adjustment for Terminal Stops

Through closer examination of the origin inference results, for some bus trips, boardings are inferred at the arrival terminal stops for a finishing trip, which is impossible in reality. A possible explanation is that sometimes bus drivers may not log off the completed trip on the iBus system on time, and thus the trip number and direction are not changed before a new bus trip begins. This means that the transactions that actually occurred at the first few bus stops on the immediately following bus trip will be inferred to have originated at the last (terminal) stop on the previous trip.

In reality, it is impossible to have passengers get onboard at the last bus stop of a trip. In order to correct this error caused by the improper use of iBus by bus drivers, the following steps are used to check for and resolve this problem:

- If on some bus trip, boardings were inferred at the last bus stop, and no boardings were inferred at the first few bus stops on this bus run's subsequent trip, then there is a problem with the iBus data at the terminal.
- In order to redistribute the incorrectly inferred boardings at the last bus stop to the correct stops, ideally, we should have a close look at the adjacent bus trips to find the ratios of boardings between the first few bus stops and the total ridership of that bus trip. Then we could apply these ratios to the bus trips with terminal problems and redistribute the boardings from the last bus stop to the subsequent trip's first few bus stops that have no boardings inferred. This method is applied to Route 70 since iBus data is missing for the first 5 stops westbound from South Kensington. However, in practice, most of the missing boardings only occurred at the first stop of the subsequent bus trip, as the operators at the iBus control center would remind drivers if they did not log into the correct trip on time. In these cases, the correction becomes much easier by simply assigning the boardings at the last stop of a route to the first stop of the subsequent trip. Even if the missing boardings occurred at the first two or three stops, it is still fine to apply this simple method as these stops can be treated as a segment, and the

load information at the segmental level still is valid for planning purposes. Moreover, it is also easier to automate this correction procedure.

In the applications to the five selected bus routes, such terminal issues are found for Routes 70 and 185. These problems are corrected before applying the origin inference results to infer destinations since the destination inference methodology using trip-chaining depends on the accuracy of the origin inference results.

3.3 Destination Inference

The process described in Chapter 2 is used here to infer passengers' destinations, implemented in a custom designed Java program that reads its inputs from an SQL database. While this program is currently separate from the origin inference program, the two could be integrated into a single program.

Since trip chaining is used to infer destinations, look-up tables are needed to find the closest stop to the next trip's boarding stop. Section 3.3.1 illustrates the procedures to generate these look-up tables.

3.3.1 Look-up Table Generation

Two look-up tables are generated based on the assumptions discussed in Section 2.1.3 for all bus stops and Underground/rail stations which are potential interchange points. These tables are used in the program to help determine the alighting stop for all bus passengers' trips. Look-up tables are derived by calculating the distances between bus stops and between bus stops and Underground/rail stations with GIS information, as demonstrated in the following steps:

(1) Select the route for analysis.

(2) Create a buffer around the selected route. The buffer represents the geographic region that is within walking distance of the selected route.

(3) Select all the bus stops within the given buffer and export them into a new layer (Layer A). These bus stops are all within walking distance of the analyzed bus route and so are potential transfer stops.

(4) Select the bus stops along the route chosen for analysis and export them into a new layer (Layer B).

(5) Create a spatial join where all points in Layer A will be given the ID of the closest stop(s) in Layer B and a distance field showing how close they are to the closest stop in Layer B. The result is a table that contains an entry for every stop in the buffer area indicating the stop in Layer B that is closest to each stop in Layer A.

(6) Export the attribute table of the spatial joint into a text file and import it into SQL. It is now ready to be used in the program.

The look-up tables consist of the following four columns which will be used to infer the destinations:

- Current Route ID (the route ID of the studied bus route)
- Direction (A or B, as routing varies by direction)
- Next boarding stop (stop in the buffer area in the GIS files)
- Nearest stop (stop on the selected bus route closest to the "next boarding stop")

The look-up table for buses is easy to generate by simply following the steps described above. For the Underground/rail, the general methodology is similar to generating the look-up tables for buses, but GIS files for the London Underground, Docklands Light Railway (DLR) and National Rail are used. The detailed techniques to generate these tables are illustrated in Appendix A.

3.3.2 Destination Inference for Each Passenger Trip

The following steps are applied:

(1) Infer origins not only for the Oyster Card users on the selected routes, but also for the users on all routes that are parallel to or intersecting the selected routes. As mentioned before, such boarding information is required when we use the "next trip" rule to infer the destinations. The origin inference rates for these related routes are around 90% of the total Oyster transactions although some routes may have lower matching rates than the five selected bus routes because they may not have as many buses equipped with the iBus system.

(2) Identify and discard Oyster Card ID/day combinations with only one trip. No destination can be inferred for these trips and these data do not enter the "seed" O-D matrix.

(3) Match each record with the next record according to the Oyster Card ID, day and time. This is the critical step in the trip chaining process.

(4) Assign the destination as the closest stop on the route being studied to the boarding stop of the subsequent transit trip.

3.3.3 Destination Inference Results

To infer the destinations for passenger trips on Route W4, we have to first infer origins for the Oyster Card users on all routes that are parallel to or intersecting W4. Of the 486511 Oyster transactions recorded for these routes, 410319 or 84% have origins inferred.

For W4, 8585 Oyster transactions are recorded, of which 5675 (66.1%) have both their origins and destinations inferred. Table 3-1 summarizes the rules for inferring the destinations for these 5675 trips and the reasons why the inference could not be completed for the remaining 2910 trips. Corresponding inference rates for Routes 70 and 185 are also included in the table. Similar results for are shown in Table A-6 of Appendix A.

The basic information about the other three selected routes is summarized here. Route 185 runs through Central London and has a much larger ridership than either Route W4 or 70. The BODS survey on Route 185 was conducted on May 12, 2009. The other two bus routes, Routes 307 and 329 are suburban connectors and the BODS survey was conducted in early June, 2009. More details are described below:

1) Route 185 runs from Lewisham to Victoria via Catford, Forest Hill, Dulwich, Camberwell Green and Vauxhall, totals 9.3 miles (15 km) long with 10 min daytime headways and 12 min headways during the early morning or late night;

2) Route 307 runs from Brimsdown Station to Galley Lane via Enfield College, Oakwood Station, High Barnet Station and Union Street, totals 10 miles (16 km) long with 10 min daytime headways and 20 min headways during the early morning or late night;

3) Route 329 runs from Turnpike Lane Bus Station to Little Park Gardens via Wood Green Station, Winchmore Hill Police Station and Saint Stephens Church, totals

5.6 miles (9 km) long with 5-6 min daytime headways and 10 min headways during the early morning or late night.

Table 3-1: Destination inference results for the BODS survey day

Reason	W4	70	185
Total transactions with destinations inferred	66.1%	62.8%	57.5%
Lack of iBus information	10%	10%	15.4%
Next boardings not in 1km buffer area	12.6%	7.1%	11.1%
Invalid OD	7.6%	9.1%	7.2%
Single trip	3.6%	4.6%	4.8%
Lack of directional information	0.1%	6.4%	4.0%
Total Oyster transactions	8585	12074	24245

For example, on Route W4, among the 8585 Oyster transactions, of the 33.9% for which an OD could not be successfully inferred:

- 3.6% are the only transactions recorded for that Oyster Card on that day (i.e., they are single trips), thus there is no "next trip" data from which to infer the destination.
- 12.6% of transactions had a next boarding stop outside the 1 km buffer surrounding Route W4, and thus the destinations could not be inferred.
- 10% are due to a lack of the iBus information necessary to infer the boarding stop for the next trip, rendering the “next trip” rule infeasible.
- 7.6% of the ODs initially inferred using these procedures result in invalid OD pairs: backward trips or trips that have the boarding and alighting inferred to be at the same stop. Among these 656 invalid passenger trips, 381 are the last transit trip of the day. Among these 381 last trips, 20 occur at terminal stops, 64 occur in the “hail-and-ride” (HR) segments, and 43 occur at loop stops. Thus, for the 656 invalid OD pairs, excluding those occurring at terminal stops, loops or HR segments, 254 invalid OD pairs are the last transit trip of the day, which account for 38.7% of the total invalid OD pairs. The “last trip of the day” rule is based on the last recorded public transport transaction of the day for an Oyster Card. However, it may not actually be the last trip of the day for each Oyster Card holder; the actual last trip of the day for those people may be by car or another travel mode, which adds more uncertainty to the modified “last trip of

day" rule. For example, someone who goes to a pub late in the evening by bus may return home via a taxi or a carpool since W4 does not run all night. More details are included in Table 3-2:

Table 3-2: Invalid OD pair analysis

Invalid OD	Trips	% of all	Last trip of the day	% of total last trips
Terminal	89	13.6%	20	22.5%
Loop	57	8.7%	43	75.4%
HR	150	22.9%	64	42.7%
Sub-total	296	45.1%	127	42.9%
Other	360	54.9%	254	70.6%
All invalid ODs	656		381	58.1%

Table 3-3 shows the origin and destination estimation results for Route W4 over two weeks from November 8 to 21, 2008:

Table 3-3: Two week OD inference results (Route W4)

Date	Total Oyster	No. of O Inferred	% of Total	No. of O& D Inferred	% of Total
Nov 8,2008, Sat	7066	6628	93.8%	3869	54.8%
Nov 9,2008, Sun	4879	4803	98.4%	2528	51.8%
Nov 10,2008, Mon	8122	7215	88.8%	3878	47.7%
Nov 11,2008, Tue	8614	8300	96.4%	4797	55.7%
Nov 12,2008, Wed	8868	7950	89.6%	4613	52.0%
Nov 13,2008, Thu	8836	7929	89.7%	4585	51.9%
Nov 14,2008, Fri	9263	6207	67.0%	3378	36.5%
Nov 15,2008, Sat	7447	6593	88.5%	4069	54.6%
Nov 16,2008, Sun	4904	4617	94.1%	2818	57.5%
Nov 17,2008, Mon	8280	8134	98.2%	5064	61.2%
Nov 18,2008, Tue	8739	7443	85.2%	4461	51.0%
Nov 19,2008, Wed	8677	8398	96.8%	5400	62.2%
Nov 20,2008, Thu	**8585**	**8212**	**95.7%**	**5394**	**62.8%**
Nov 21,2008, Fri	8631	7549	87.5%	4521	52.4%
Total/Mean	**110911**	**99978**	**90.1%**	**59375**	**53.5%**

In general, for Route W4, for most days, more than 50% of trips have both origin and destination inferred. On Nov 20, 2008, when the BODS survey was conducted, 95.7% of the Oyster transactions on Route W4 have origins inferred and 62.8% of the Oyster transactions also have destination inferred. Similar analyses for the other selected routes summarized in Table 3-4 show that the two-week OD inference rates are quite close to the survey days.

Table 3-4: Origin and destination inference results*

Bus Routes	No. of Oyster Transactions	No. of Origin Inferred	% of Origin Inferred	No. of Destination Inferred	% of Destination Inferred
W4	8585	8212	95.7%	5393	62.8%
70	12074	11381	94.3%	7741	64.1%
185	24245	22794	94.0%	13947	57.5%
307	10057	9456	94.0%	6968	69.3%
329	17496	17033	97.4%	13737	78.5%

* These results are for the BODS survey days.

3.4 Summary

The OD matrices provide fundamental information for public transit planning and operations analysis. This chapter takes a first step towards the integration of the archived AFCS and AVL data for transit planning in London. The algorithm takes advantage of the pattern of a passenger's consecutive transit trips and uses the next trip's boarding location information to infer the destination of the prior transit trip. The London analysis examines the bus-to-bus and bus-to-Underground/rail trips by matching the AFC system against the iBus AVL system, and examines the spatial relationship between the bus and Underground/rail networks using GIS technology.

The inference process has been shown to work fairly well. As shown in Figure 3-8, of all the bus passenger trips using the Oyster Cards on the five selected routes, more than 90% of the bus passenger trips have the origins inferred and more than 50% of the bus passenger trips have both the origins and destinations inferred. Such results can provide very useful statistics regarding the demonstrated demand for the bus service provided by transit operators.

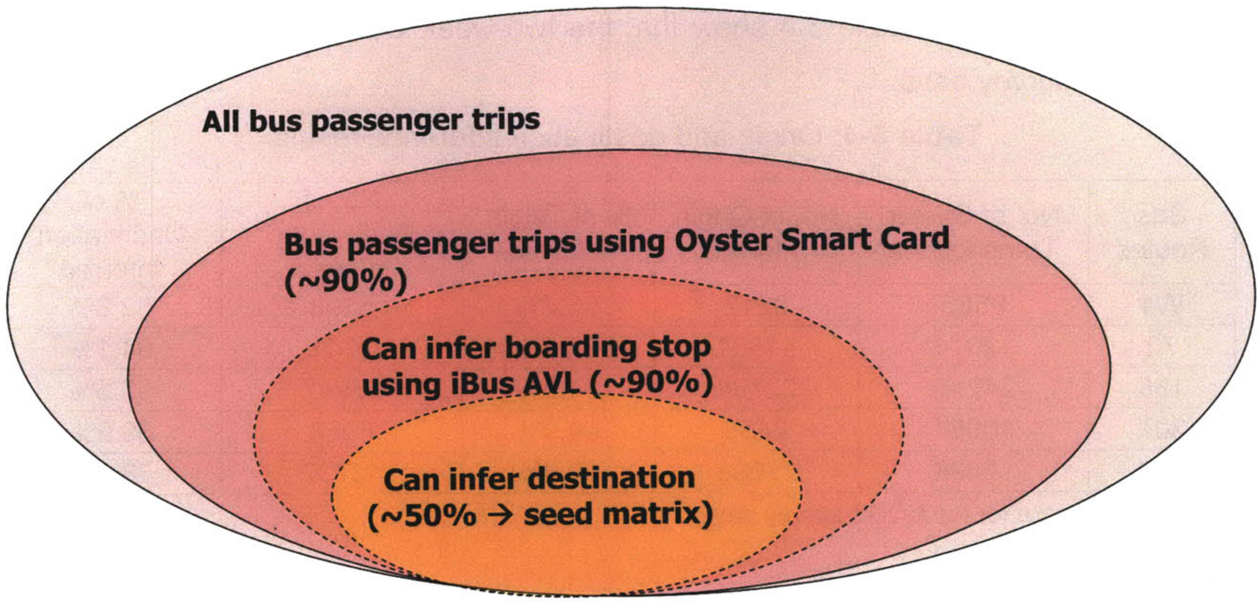


Figure 3-8: Relationship between inferred "seed" OD matrix and the original data

Of course, there are limitations to this methodology. For example, the usage of the Oyster Cards is not universal in London. As shown in Figure 3-8, Oyster transactions account for about 90% of all bus passenger trips in London while the origins and destinations for the other 10% of bus passenger trips cannot be inferred using this methodology. Even for the 90% of all bus passenger trips that use the Oyster Cards, it is not possible to infer both origins and destinations for all because of a lack of the iBus AVL data or other missing information. However, based on the available data sources, the results are fairly good, and even a 50% valid sample is a huge increase in the available data upon which to base future planning decisions. Chapter 4 will further validate the inferred origins and destinations by comparing the results with the surveyed bus trips in the BODS dataset.

4 Origin-Destination Inference Validation

A thorough evaluation of the methodology for inferring the origins and destinations as well as a validation of the resulting inferred OD matrices is necessary before they can be applied in practice. Two criteria are critical regarding the OD estimates. First is the robustness of the proposed methodology; and second is, the proximity of the inferred origins and destinations to the origins and destinations obtained from the periodic one-day BODS on-board passenger surveys.

This chapter validates the inferred origins and destinations for the selected bus routes in London. It begins with a primary analysis by comparing the ridership totals from the BODS survey, ETM data and Oyster transactions for all surveyed bus trips. Section 4.2 then compares the BODS and Oyster transactions for the unexpanded data; specially, it compares the origins inferred from Oyster transactions with the BODS survey results for all the surveyed trips. Next, Section 4.3 compares the BODS surveyed destinations and the results from the Oyster inference methodology. Section 4.4 summarizes the results of the validation.

4.1 Comparison of BODS, ETM and Oyster Data for all Surveyed Trips

To validate the inference results for the origins and destinations, three data sources are used: the BODS survey data, ETM data for each bus trip and Oyster transactions. Since these datasets come from different systems located in different departments within TfL and are collected in different ways, it is possible that these datasets may be inconsistent with each other. This section explores the unexpanded (surveyed trips) BODS, ETM data and Oyster transactions to check for data consistency.

4.1.1 Aggregate Comparison between BODS, ETM and Oyster for all Surveyed Trips

This section compares the ridership estimates from BODS, ETM data and Oyster transactions for all surveyed trips. Ideally, for all these trips, the aggregate BODS-issued counts should be equal to the ETM counts. Since Oyster transactions record only passengers using Oyster Cards, the BODS-issued counts should be no less than the Oyster counts.

One issue with the BODS survey data is that not all issued survey cards are returned. This analysis for the BODS data uses the numbers of both issued and returned survey cards and examines how the total ridership varies between BODS, ETM and Oyster data for all surveyed bus trips (see Table 4-1 and Figure 4-1).

Table 4-1: Aggregate ridership comparison between BODS, ETM and Oyster Datasets

Categories	W4 (108 bus trips)	70 (112 bus trips)	185 (138 bus trips)	307 (116 bus trips)	329 (214 bus trips)
BODS Issued	5417	8161	17833	7313	12410
BODS Returned	5035	7560	14650	6504	10489
ETM Total	5675	8513	18455	7287	12935
Oyster	5400	8054	16983	6797	11992

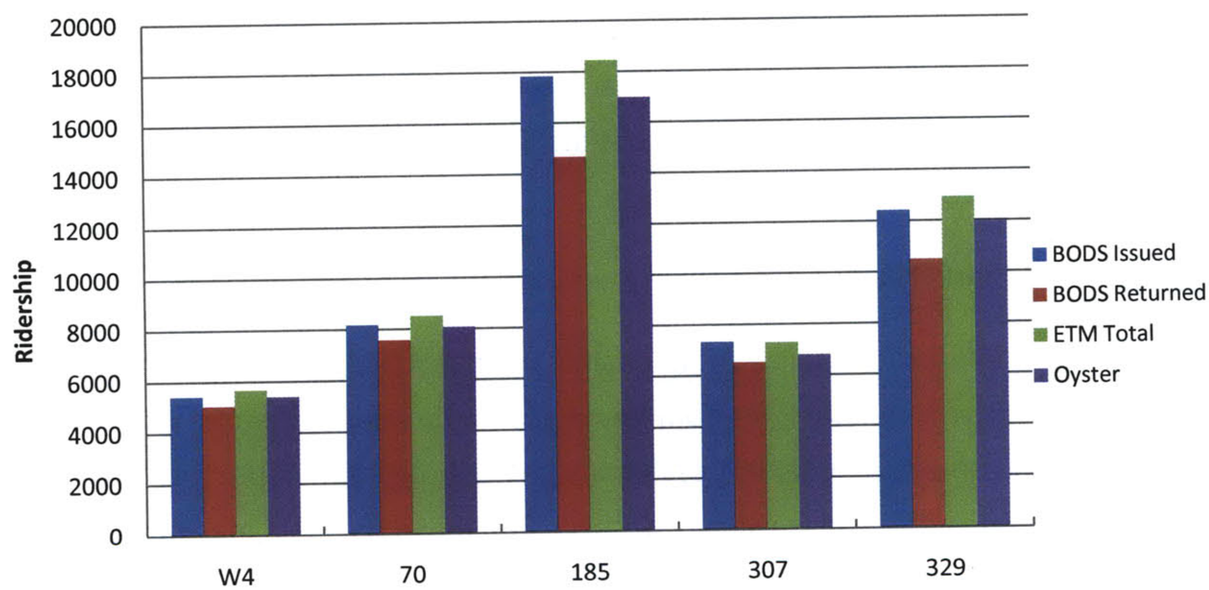


Figure 4-1: Aggregate ridership comparison between BODS, ETM and Oyster

Routes W4, 70 and 307 are similar with regard to the ridership comparison among BODS, ETM and Oyster datasets. The completed (returned) BODS counts are about 93% of the BODS-issued counts, and the BODS-issued counts are 4.5% (or 258) less than the ETM total on Route W4 and 4.1% (or 352) less on Route 70. However, Route 185 is different, with much lower return rates (82%) for the survey cards. Since Route 185 runs through Central London and has much higher daily ridership than the other four routes, getting a high return rate might be difficult, especially during peak hours. The return rate for Route 329 is also low (84%). Though it is a suburban connector and the shortest route (around 9 km) selected, the headway of Route 329 is the shortest, 5-6 minutes in daytime and 10 minutes during early morning or late night. Thus, there are more bus trips running due to high demand and being surveyed during the peak periods, and crowding may be the cause of lower return rates on this route.

In general, the "BODS-Issued" numbers are 3-4% lower than the "ETM Total" and are slightly higher than the "Oyster" numbers. Three possible explanations exist. It could be that BODS surveyors do not issue a card to every boarding passenger, and thus the BODS counts are lower than "reality". Or it could be that the ETM totals overestimate the total loadings for some bus trips. Finally, it could be that the Oyster passengers make up a greater proportion of total passengers than thought. In order to get insight into which of these options best explain the ridership differences, Section 4.1.2 analyzes all the BODS surveyed bus trips in detail.

4.1.2 Trip Level Comparison between BODS, ETM and Oyster

This section compares the bus-trip level ridership from BODS, ETM and Oyster datasets for each surveyed trip in order to get a clear picture of where the differences in the above aggregate analysis arise. Note that the BODS counts used here are the number of survey cards distributed (even though some were not returned) and the comparisons include all the BODS surveyed trips. In order to compare the data at a detailed level, the day is divided into five time periods: Early AM (4:30-6:59); AM Peak (7:00-9:29); Midday (9:30-15:59); PM Peak (16:00-18:29) and Late (18:30-4:29 the next day).

The comparison results for Route 329 are presented in Figure 4-2, with similar results for Routes W4 and 70 shown in Appendix B as corroborative evidence. Route 185 presents a different picture, which is partly due to the fact that it runs through the Central London area and is one of the busiest routes in London. More details about this route are attached in Appendix B.

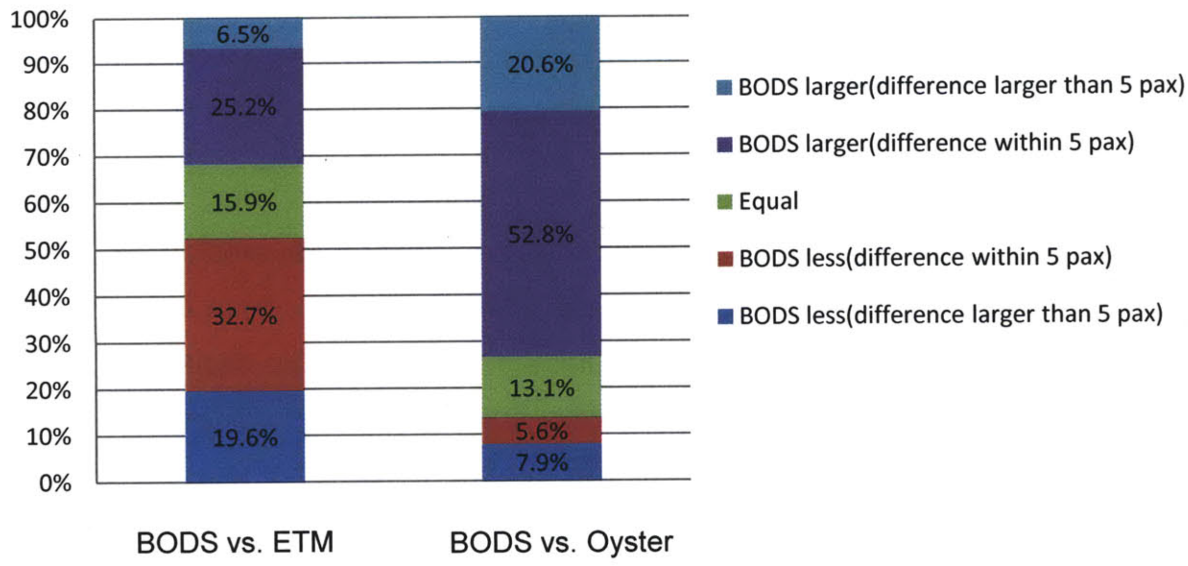


Figure 4-2: BODS-ETM-Oyster ridership comparison for Route 329

Figure 4-2 shows the comparisons between BODS and ETM, and BODS and Oyster for Route 329. For 15.9% (or 34 trips) of the 214 surveyed bus trips, the BODS and ETM counts are identical. The ridership difference between BODS and ETM for a further 124 trips (57.9%) is less than 5 passengers. These three categories of bus trips (abs (BODS-ETM) ≤ 5 passengers) include 7990 passengers (or 64.4% of all BODS surveyed passengers).

However, there are some trips for which the results are harder to explain. The BODS counts are much smaller than the ETM data for 42 bus trips (19.6%), with an average of 15 fewer passengers per trip. All these trips occurred during the PM Peak or Late time period with much lower survey sample return rates. Furthermore, the BODS counts are also lower than the Oyster counts for most of these 42 trips. It seems likely that the BODS survey severely undercounts the ridership for these trips, especially at peak hours or late at night.

For another 14 bus trips (6.5%), the BODS counts are much larger than the ETM counts, with an average of 11 more passengers per trip. For most of these trips, the ETM counts are almost the same as the Oyster counts, which implies that there are no non-Oyster passengers since the ETM counts are expected to count all boarding passengers regardless of whether they use the Oyster Cards or not. It is quite possible that the ETM counts underestimate the ridership for these trips. Since the way ETM records non-Oyster passengers is by bus drivers manually pressing a button when a non-Oyster passenger boards, it is likely that some bus drivers do not push the button every time a non-Oyster passenger boards. Thus the ETM counts may not capture all the non-Oyster passengers for these particular trips.

The BODS counts are equal to, or larger than, the Oyster counts for 86.4% of all bus trips (or 185 trips). However, there are 17 trips (29.6%) for which the ridership from the BODS is much lower than that from the Oyster, with an average of 18 fewer passengers per trip. The BODS ridership for 12 of these 17 trips is much lower than the Oyster ridership (the differences in ridership are greater than 10 passengers per trip). Most of these trips occur in the same direction (Northbound on Route 329). As mentioned above, it is quite possible that the BODS counts underestimate ridership for these 12 trips.

As shown in Figure 4-2, in general, for 73.8% of all the surveyed bus trips, the ridership in the BODS dataset is equal to or close to that in the ETM dataset (the differences in ridership ≤ 5 passengers). For those unexpected cases (e.g., the ridership from the BODS dataset is much lower than that from the ETM dataset), some seem to be caused by the BODS dataset underestimating the number of passengers since the BODS counts are also smaller than the Oyster counts for these trips. For the other unexpected cases (e.g., the ridership from the BODS dataset larger than that from the ETM dataset), some seem to be caused by the ETM counts underestimating the number of non-Oyster passengers since the ETM counts do not record any non-Oyster passengers for several bus trips. With regard to the comparison between the BODS and Oyster datasets for each surveyed trip, most of them meet our expectation. The few exceptions are when the ridership from BODS is much lower than that from the Oyster dataset (7.9%). The explanation is likely to be that the BODS counts underestimate the ridership for some trips.

4.1.3 Summary

Based on the ridership comparison results for these surveyed trips, data inconsistencies appear among these three datasets (BODS, ETM and Oyster), which will help explain some of the differences between the surveyed O-D flows and the automatic data inferred O-D matrices. The possible sources of error that might cause these data inconsistencies include: 1) BODS may underestimate the ridership in peak hours or lunch time; 2) when BODS survey return rates are lower than usual (50%-80%), there is a possibility that not all the passengers were issued a survey card; 3) ETM may underestimate the ridership (drivers may not push the button for each boarding non-Oyster passenger on specific trips); and 4) when the ridership from the ETM dataset equals to or is slightly larger than that from the Oyster dataset, the Oyster passengers may make up a greater proportion of total passengers than we would expect or the ETM data may be missing some non-Oyster passengers.

However, the majority of the data are consistent among the three datasets. Hence, it is reasonable to expect that the inferred aggregate OD matrices using the Oyster dataset should closely approximate the actual bus passengers' travel patterns. Sections 4.2 and 4.3 validate the accuracy of the inferred origins and destinations, respectively.

4.2 Comparison of Boarding Locations between BODS and Oyster Datasets

This section compares the results of the origin inference method with the boarding locations from the BODS survey. As shown in Figure 4-1, the number BODS returned cards accounts for around 94-96% of the number of Oyster transactions for the surveyed trips, with the exception of Routes 185 and 329, where the percentage is 86% and 87%, respectively. Moreover, in the BODS database, only returned BODS cards are counted in the OD matrices. Therefore, though non-Oyster passengers are not included in the automatic OD inferences and the origin inference rates are around 94% for all the selected routes as shown in Table 4-2, the total number of boardings inferred from Oyster transactions might be larger or equal to that from the BODS surveyed OD matrices for all the surveyed trips.

Table 4-2: Origin inference rates for the selected routes on their survey days

Bus Routes	No. of Oyster Transactions	No. of Origin Inferred	% of Origin Inferred
W4	8585	8212	95.7%
70	12074	11381	94.3%
185	24245	22794	94.0%
307	10057	9456	94.0%
329	17496	17033	97.4%

Since the origin inference rates are quite high and, the BODS surveyors do not receive return surveys from all passengers (the sample rates for some bus trips are as low as 60%), the total number of boardings inferred from the Oyster transactions is close to that from the BODS survey. Consequently, the number of boardings at each stop from the Oyster estimates should be close to that from the BODS database if the origin inference method works well.

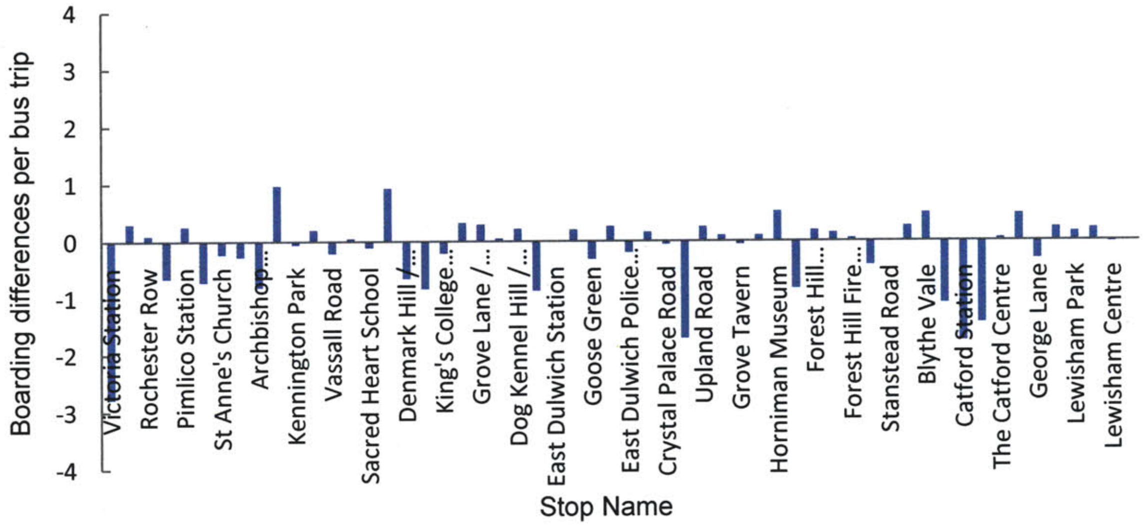
Routes 185 and 307 are chosen here as examples to show the comparison results at the bus-stop level. Tables 4-3 and 4-4 summarize the total number of boardings from BODS and Oyster datasets by direction on Routes 185 and 307 for the surveyed bus trips, respectively. For Route 185, which is one of the busiest bus routes in London, the BODS survey return rate is the lowest of the five selected routes (86%), and the number of boardings from BODS is significantly lower than that from the Oyster transactions for the surveyed trips, as shown in Table 4-3. For Route 307, which is a suburban connector, the situation is very different. The BODS return rate is the highest among the five selected routes (97%), and the number of boardings from BODS is slightly higher than that from the Oyster transactions for the surveyed trips, as shown in Table 4-4. Figures 4-3 and 4-4 further demonstrate the comparison results at the bus-stop level for these two routes that have the lowest and highest sample return rates, respectively.

Table 4-3: Summary of boardings from BODS and Oyster (Route 185)

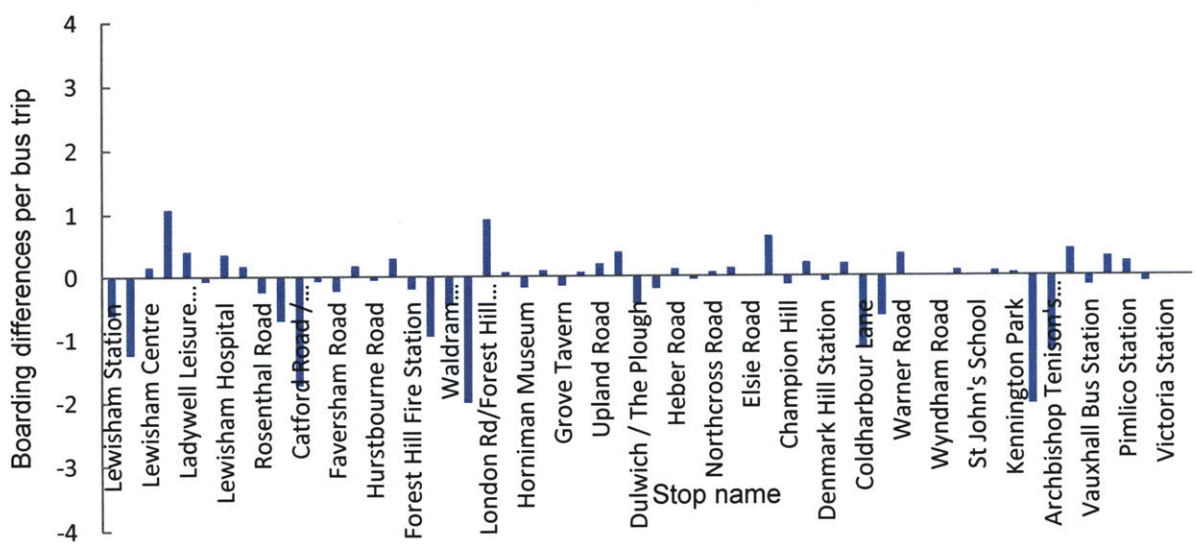
Direction	No. of BODS boardings	No. of Oyster boardings	No. of surveyed bus trips
Eastbound	7304	7911	66
Westbound	6904	7386	62

Table 4-4: Summary of boardings from BODS and Oyster datasets (Route 307)

Direction	No. of BODS boardings	No. of Oyster boardings	No. of surveyed bus trips
Eastbound	3185	3169	56
Westbound	3296	3221	58



(a) Eastbound

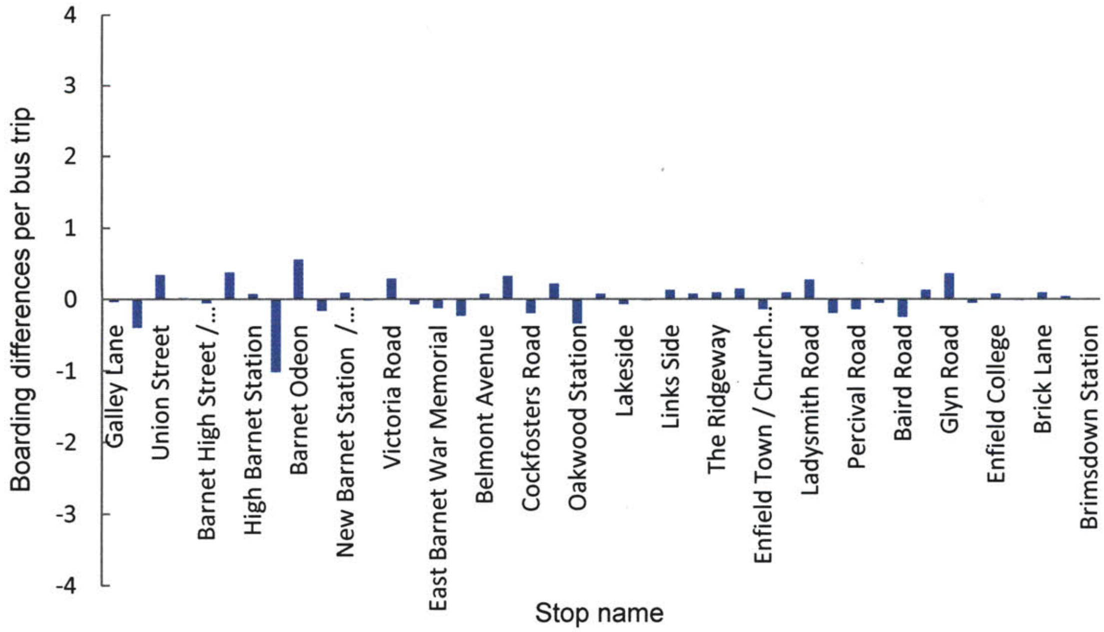


(b) (Westbound)

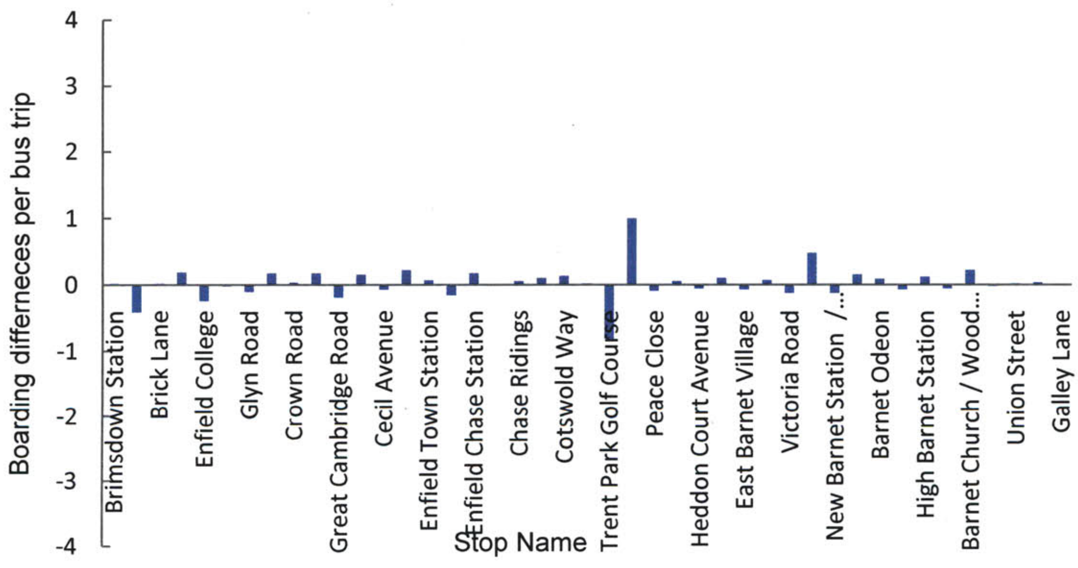
Figure 4-3: Boarding location comparison for Route 185

The total number of boardings for the surveyed trips from BODS is 607 passengers fewer than that recorded in the Oyster dataset on Route 185 eastbound, and 482 passengers fewer than Oyster in the westbound direction. These relatively large boarding differences between BODS and Oyster datasets are mainly due to the low BODS survey sample return rates, because Route 185 is one of the busiest routes in London. The boarding differences between BODS and Oyster at some stops are larger than other routes (see Figures B-5 and B-6 for Routes W4 and 70 in Appendix B), up to 2 or 3 passengers per trip (usually with the Oyster passenger trips being greater than BODS) as shown in Figure 4-3, while it is usually about 1 passenger per trip at each stop on other routes. Most of these stops where the boarding differences are larger than 1 passenger per trip happen to be close to shopping centers or problem stops as listed in the BODS report from TfL. One of the stops that has the largest boarding differences on Route 185 is Victoria Station, where BODS has about 3 fewer passengers boarding than the Oyster dataset. This big difference is actually caused by one bus trip (No. 140 during the PM Peak time period), when both ETM and Oyster ridership are much larger than BODS. Meanwhile, the BODS return rate for this trip is 65% as indicated in the BODS report. It seems likely that the BODS surveyors did not issue the survey cards to all the boarding passengers in this case.

However, Route 307 is totally different from Route 185 as it is a suburban connector with a high sample return rate (about 97% of the BODS survey cards are returned). The number of boardings from the BODS survey is almost the same as that from Oyster (see Table 4-4). Hence, the number of boardings from BODS per trip at each stop is expected to be the very close to that from the Oyster. Figure 4-4 supports this expectation for both directions on Route 307.



(a) Eastbound



(b) Westbound

Figure 4-4: Boarding location comparison for Route 307

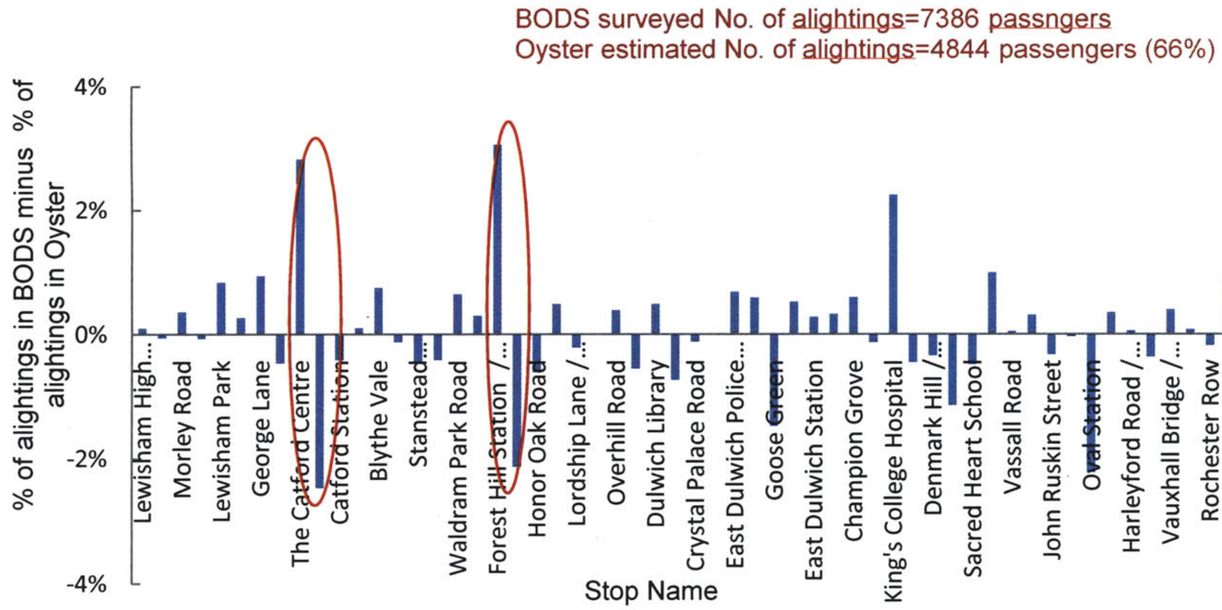
The boarding location comparison results for the other three selected routes support the same conclusion that in general, the average number of boardings per stop from the Oyster estimates is close to that from the BODS survey for all of the BODS surveyed trips. The only significant differences seem to be caused by the low BODS sample rate. Overall, the Oyster origin inference methodology works well and thus could be used to further infer the bus passengers' destinations as well as to provide more comprehensive, reliable and cost-effective information for transit planners.

4.3 Comparison of Alighting Locations between BODS and Oyster Datasets

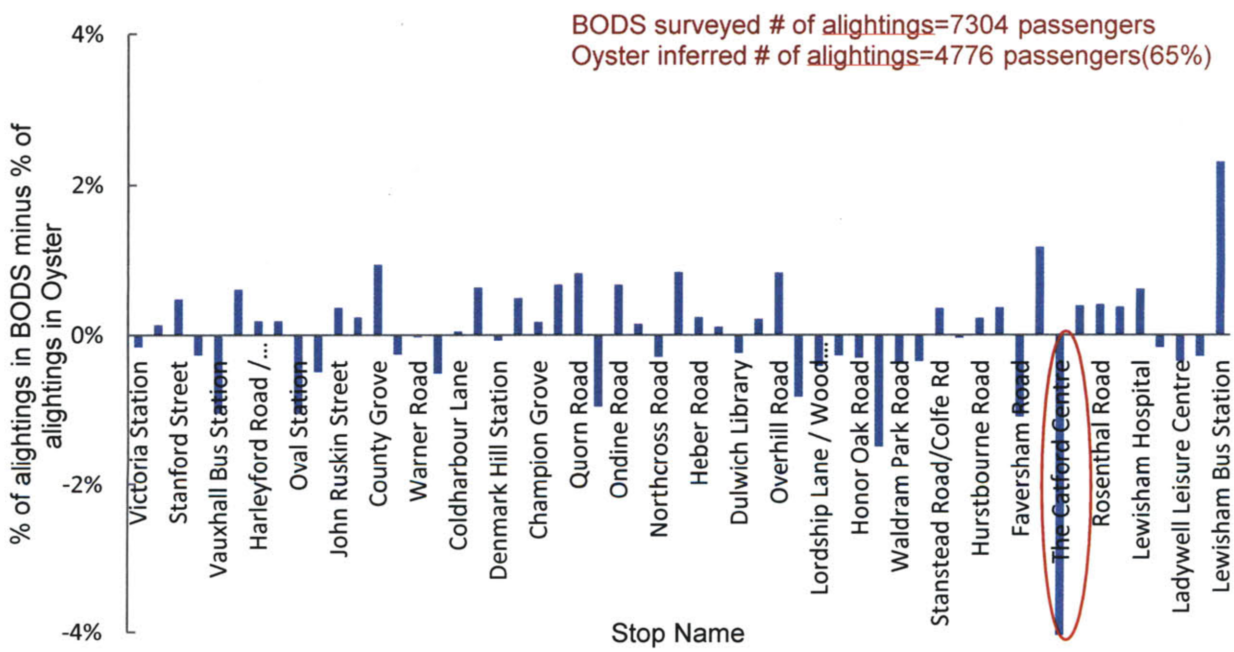
This section tests the destination inference method by comparing the percentages of alightings at each stop in the BODS dataset with the Oyster estimates. Since destinations could be inferred for only about 60% (see Table 4-5) of all the Oyster transactions on the selected routes, the number of inferred alightings at each stop from the Oyster estimates will typically be far less than the BODS survey results. But we expect the percentages of inferred alightings from the Oyster estimates to be close to the percentages of alightings from the BODS for all the surveyed bus trips. Figure 4-5 demonstrates the results of comparing the alighting locations in the case of Route 185.

Table 4-5: Destination inference rates for the selected routes on their survey days

Bus Routes	No. of Oyster Transactions	No. of Destination Inferred	% of Destination Inferred
W4	8585	5393	62.8%
70	12074	7741	64.1%
185	24245	13947	57.5%
307	10057	6968	69.3%
329	17496	13737	78.5%



(a) Northbound



(b) Southbound

Figure 4-5: Alighting location comparison for Route 185

As shown in the above figures, 7386 alightings were recorded in the BODS survey, while 4844 Oyster passengers (66% of BODS surveyed alightings) have their destinations inferred on Route 185 northbound. For the southbound direction, 4776 Oyster passengers have their destination inferred (65% of BODS surveyed alightings). For both directions, the number of inferred Oyster alightings is far lower than the BODS survey results. However, the percentage of inferred alightings from the Oyster estimates at each stop is very close to the percentage of alightings from BODS for all the surveyed trips, with the differences generally within 2%. There is a relatively large difference (4%) between the BODS dataset and the Oyster estimates at the Catford Shopping Center bus stop on Route 185 southbound. The BODS validation report mentions some problems here as several issued cards were not returned, which caused part of the differences here. The other reason is that passengers might not necessarily get off the bus at the stop that is closest to their next boarding stop, especially when the stops are close to a shopping center, where people may walk around more than usual. Another large difference appears at Victoria Station on Route 185 northbound, which has connections to five other bus routes, the Underground and National Rail. It is quite possible that the BODS survey cannot reach all the passengers at this bus stop due to the crowding. However, for most of the other stops, if the percentage of alightings in the BODS survey differs greatly from that in the Oyster estimates, these differences are generally offset at adjacent stops, as shown by the red circles in Figure 4-5. As mentioned above, passengers might get off the bus one stop before if their desired destinations are almost equidistant from a stop or if they have several nearby errands to complete, which our model cannot capture. The comparison results for other selected routes attached in Appendix B support the same conclusion.

In general, the percentage of alightings at each bus stop from the Oyster estimates closely matches the percentage from the BODS survey. For the larger differences observed at certain bus stops, most of those differences between the BODS dataset and the Oyster estimates can be offset with opposite differences at adjacent stops. Though the method assumes that passengers will alight from buses at the closest stop to the next boarding location, this will not be universally true as some passengers may want to walk short distances, especially in business and commercial areas. Therefore, although we cannot infer the destinations perfectly for all the trips, these comparison results show that the destination inference methodology should provide transit planners generally reliable information on the spatial and temporal distribution of the alightings, at least at the route-segment level.

4.4 Summary

This chapter validates the OD inference results in Chapter 3 by first comparing the total ridership in the BODS survey and Oyster transactions for all the BODS surveyed trips. Next it compares the boarding and alighting locations separately, using the BODS survey and the Oyster estimates for each bus stop. For the ridership comparison, results show that most of the data are consistent among the three data sources used in this research, namely, the BODS, ETM dataset and Oyster transactions, though there are some inconsistencies for certain bus trips. For the boarding and alighting locations at each bus stop, the OD inference results are very similar to the BODS survey results. There are some small bus-stop level differences between the Oyster and BODS results, but these are either caused by low survey sample rates or offset at adjacent bus stops. Thus the inferred ODs are highly consistent with the BODS ODs at the segment level.

To sum up, the automatic inference process has proven to work well when compared to the BODS survey. The next chapter will discuss potential applications of the results of these origin and destination inferences.

5 Application of the Inferred Origin-Destination Matrices

The validations conducted in Chapter 4 have shown that the origin and destination inference process using the proposed methodology works well when compared with the BODS survey results. This chapter presents several applications using the results of the inference process. One of the most significant applications is using the automatic OD inference to better understand bus passengers' travel patterns on a daily basis. Surveys are limited by narrow spatial and time coverage, while an automatic procedure can generate OD matrices for any bus route at any time at a low marginal cost, as long as ADC systems are deployed and procedures developed. This chapter begins with a characterization of the daily ridership variation for each bus trip on selected routes over a two-week period. Section 5.2 further illustrates the daily variations by showing the load profiles for both weekdays and weekends. Section 5.3 assesses the trip distance distributions. Section 5.4 demonstrates another application of the inferred OD matrices by analyzing the interchange times for bus passengers. Section 5.5 analyzes the subsequent travel behaviors for bus passengers whose observed trips are on selected routes. The chapter ends with a summary of how these applications can contribute to improved transportation planning in London.

5.1 Daily Ridership Variation

This section explores the daily ridership variation for each bus trip over two weeks based on the Oyster transaction data. Before the detailed analysis for the trip level ridership, the daily ridership variations between Oyster and non-Oyster passengers over two weeks for the five selected routes are shown in the following section.

5.1.1 Oyster and non-Oyster ridership based on ETM data

The ETM data can provide a good idea of daily ridership variation for Oyster and non-Oyster bus passengers. Two weeks of ETM data for the five selected routes are examined to show this variation.

- Route W4

For Route W4, Oyster passengers accounted for 90-96% of the daily ridership, except on Saturday, November 22, 2008, when Oyster passengers accounted for only 82% of the day's ridership. The average weekday ridership is 9575, with an average Oyster ridership of 8828, 92% of the total.

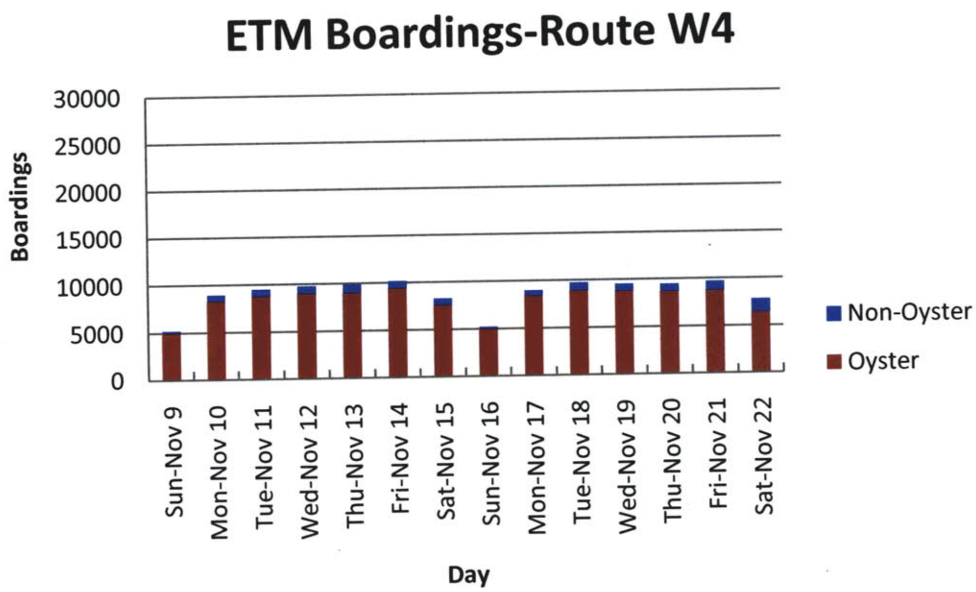


Figure 5-1: Daily ridership variation for Route W4

- Route 70

For Route 70, Oyster passengers accounted for 92-95% of the daily ridership, except on Saturday, November 22, 2008, when Oyster passengers accounted for 88% of the day's ridership. The average weekday ridership is 12788, with an average Oyster ridership of 12092, 94% of the total.

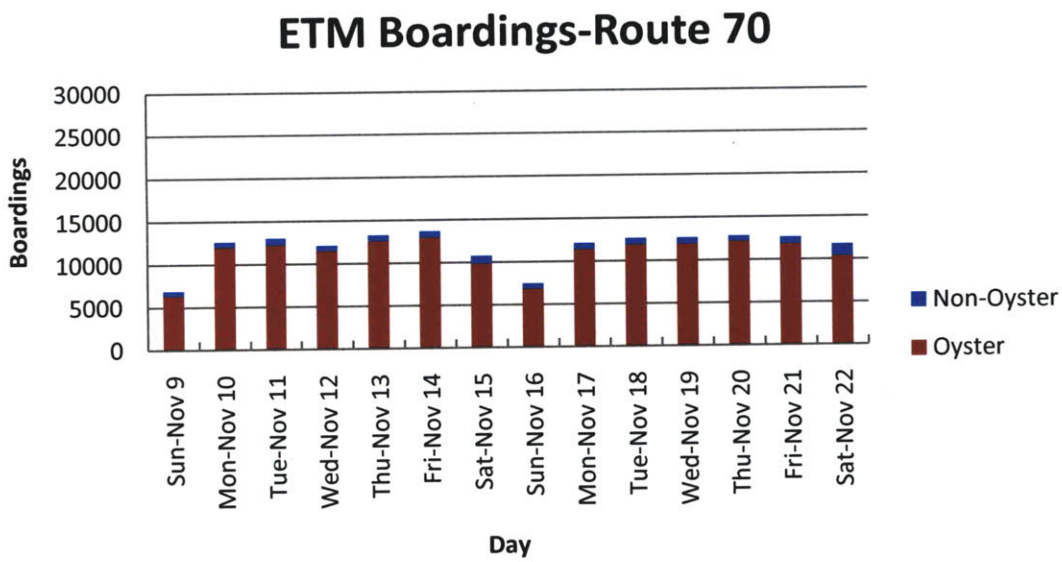


Figure 5-2: Daily ridership variation for Route 70

- Route 185

For Route 185, Oyster passengers accounted for 91-93% of the daily ridership, except on Sunday, June 7, 2009, when Oyster passengers accounted for 88% of the day's ridership. Weekdays tend to have a higher percentage of Oyster passengers. The average weekday ridership is 26030, with an average Oyster ridership of 24102, 93% of the total.

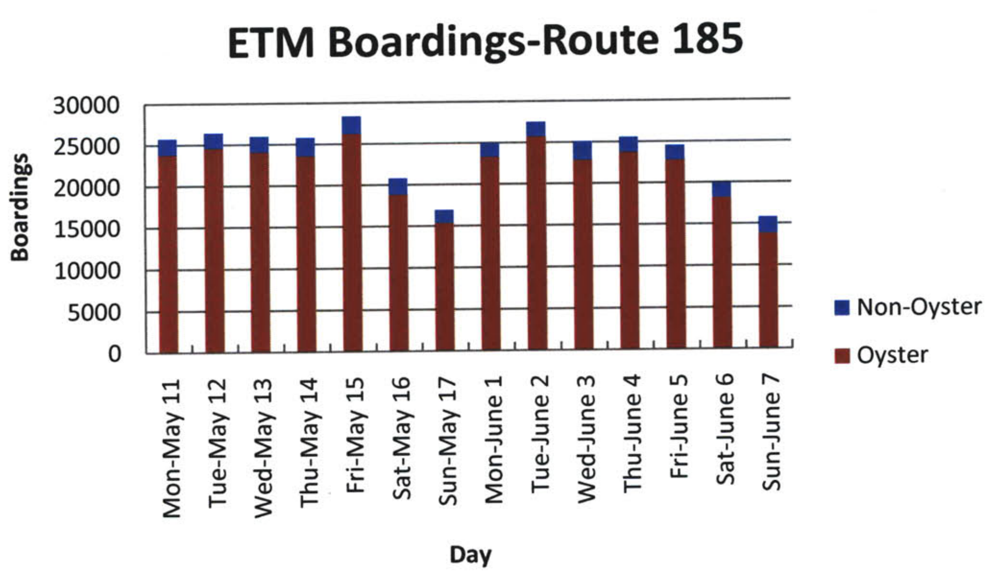


Figure 5-3: Daily ridership variation for Route 185

- Route 307

For Route 307, Oyster passengers accounted for 92-95% of the daily ridership, except on Sunday, June 7, 2009, when Oyster passengers accounted for 90% of the day's ridership. The average weekday ridership is 11619, with an average Oyster ridership of 10902, 94% of the total.

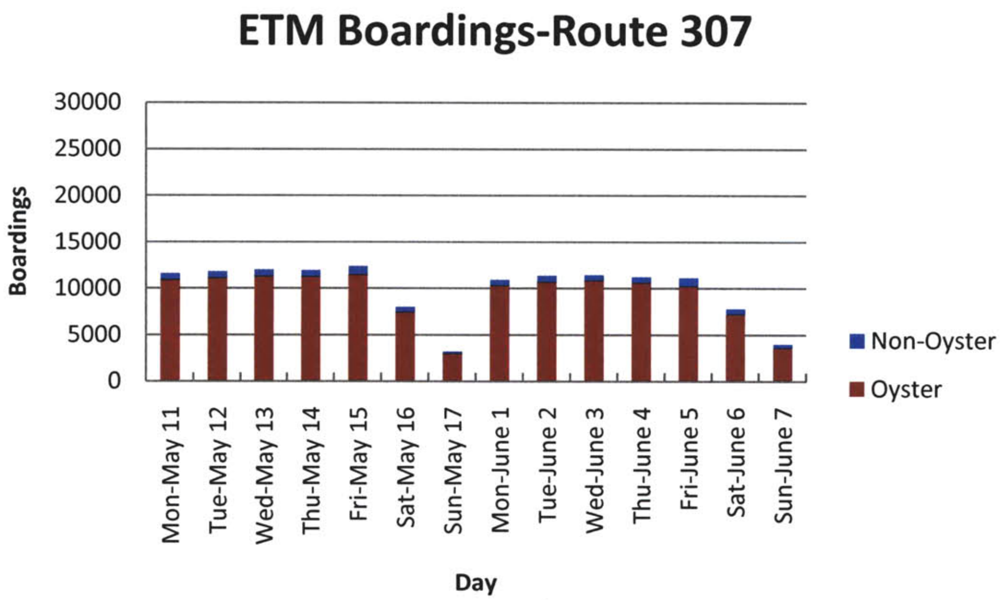


Figure 5-4: Daily ridership variation for Route 307

- Route 329

For Route 329, Oyster passengers accounted for 93-96% of the daily ridership. The average weekday ridership is 19337, with an average Oyster ridership of 18176, 94% of the total.

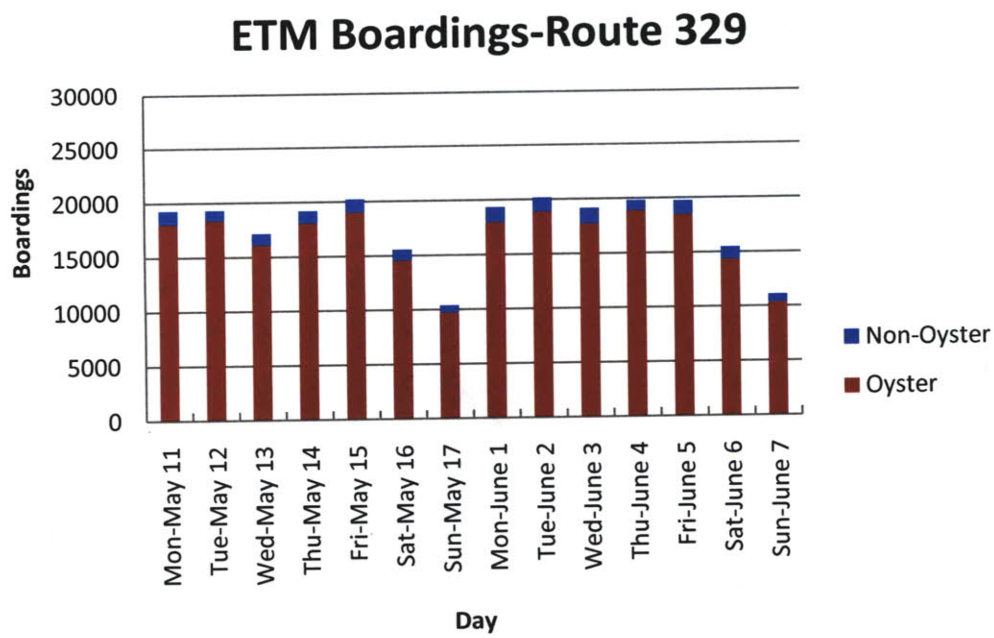
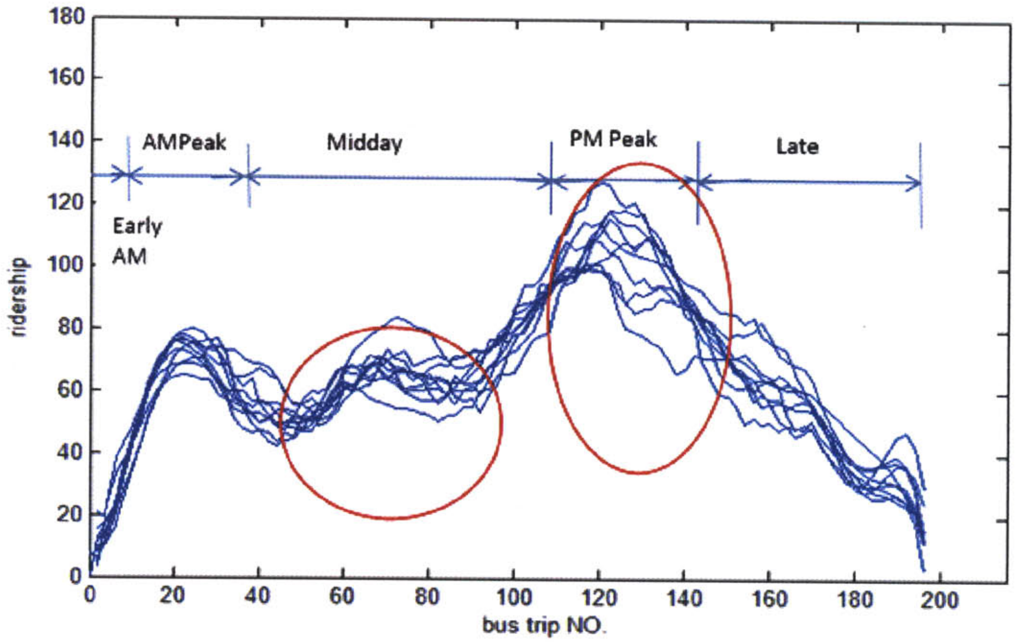


Figure 5-5: Daily ridership variation for Route 329

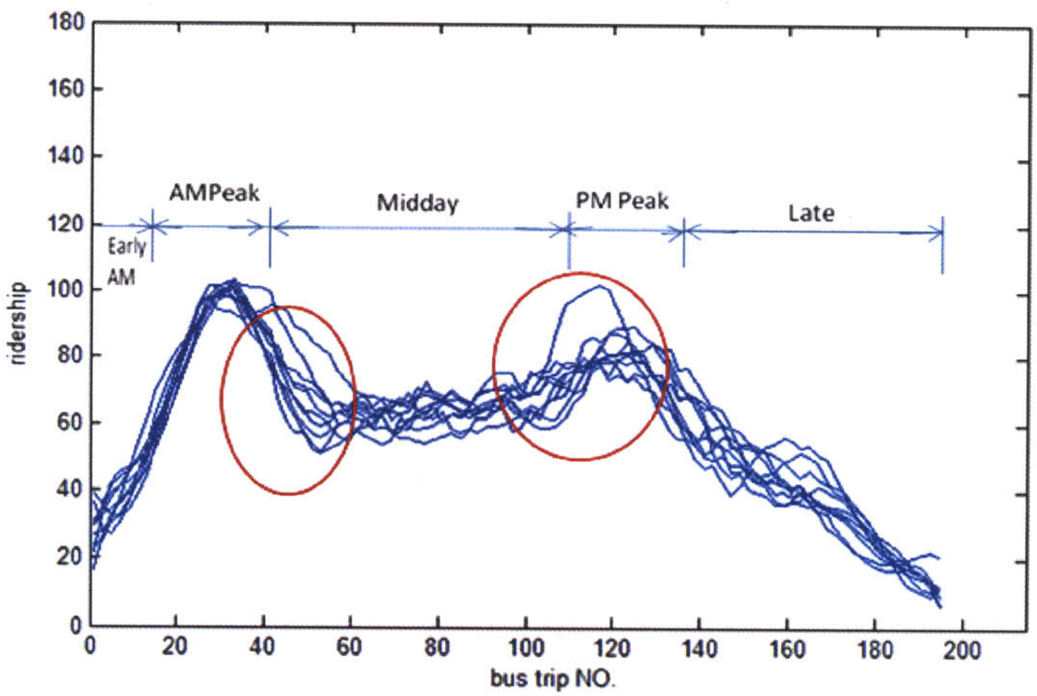
In general, for the five selected routes, the daily ridership did not vary greatly across weekdays, and Oyster passengers account for more than 90% of all bus passengers. But there are large differences in ridership between weekdays and weekends, and the percentage of Oyster passengers often varies between weekdays and weekend. Such information is important for bus service planning and fare policy, and it is easy to obtain such information from the archived ETM data for any day and any route.

5.1.2 Daily ridership variations for each bus trip

Not only does the total ridership vary daily at the route-level, the ridership for each bus trip can also vary across days. Routes 70 and 185 are chosen here to investigate this ridership variation at the trip-level, using the recorded Oyster transaction data. Figure 5-6 shows the weekday ridership variation for each bus trip on Route 70 over ten weekdays (Nov 10-21, 2008). Each line represents the trip-level ridership on a specific weekday. The x-axis displays the time of each bus trip and the y-axis displays the corresponding ridership. The red circles (for bus trips in the Midday and the PM Peak periods, respectively) show that there can be very large variances in the trip-level ridership across weekdays.



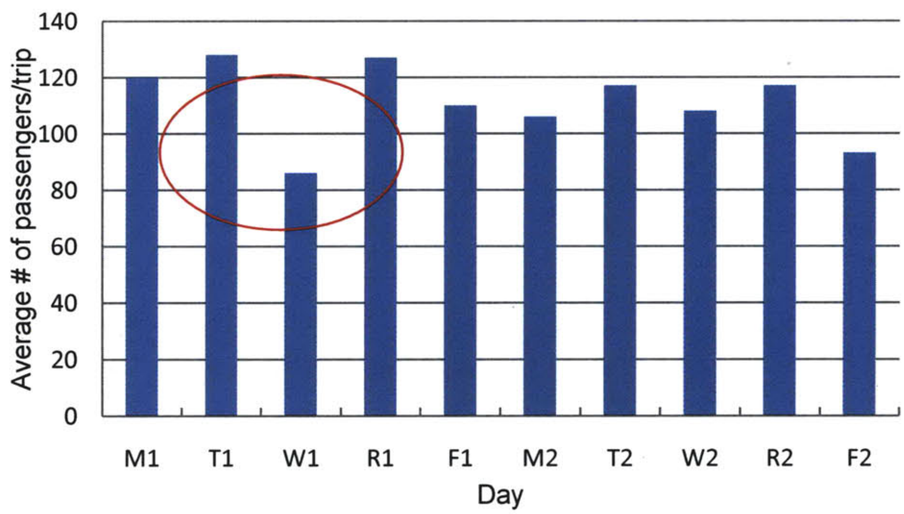
(a) Eastbound



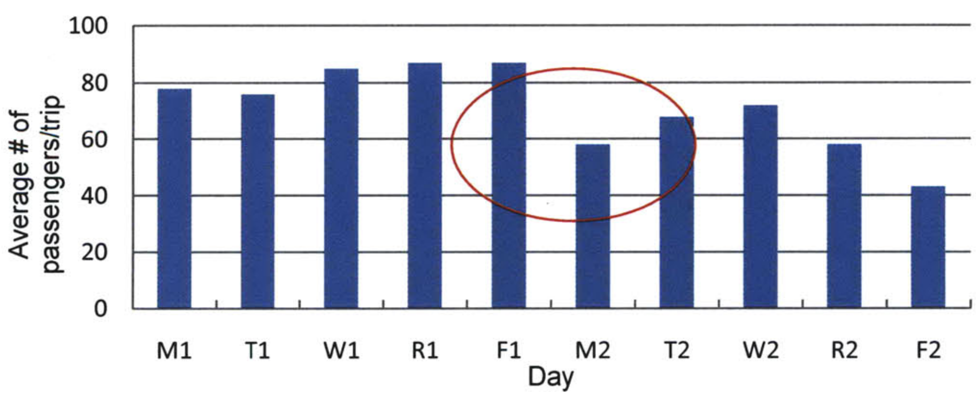
(b) Westbound

Figure 5-6: Daily variation in trip-level ridership for Route 70

Six PM Peak bus trips are selected for more detailed analysis. In order to minimize the effects of bus bunching and service gaps, the average number of passengers for three successive bus trips in the same direction is used instead of the trip-level ridership. Figure 5-7 shows the variation of the average number of passengers for three successive bus trips during the PM Peak. As shown by the red circles, large ridership differences exist for these bus trips across the studied ten weekdays. Thus any one-day survey (such as BODS) cannot accurately represent the average weekday over a period of several weeks, let alone over a period of months or years.



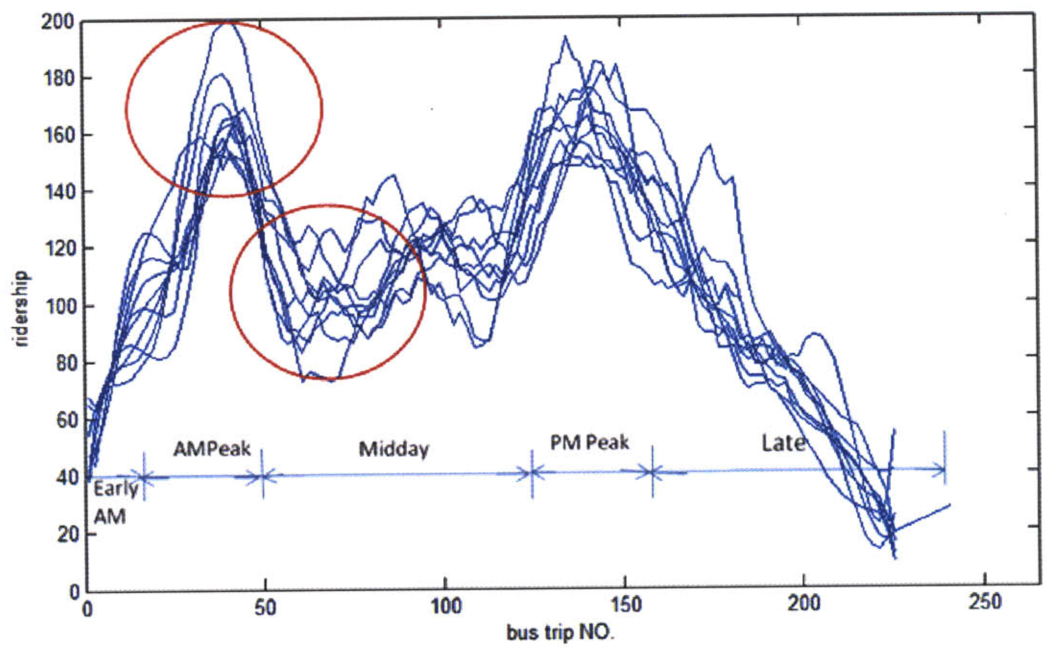
(a) Average of Bus Trips 122, 124, 126



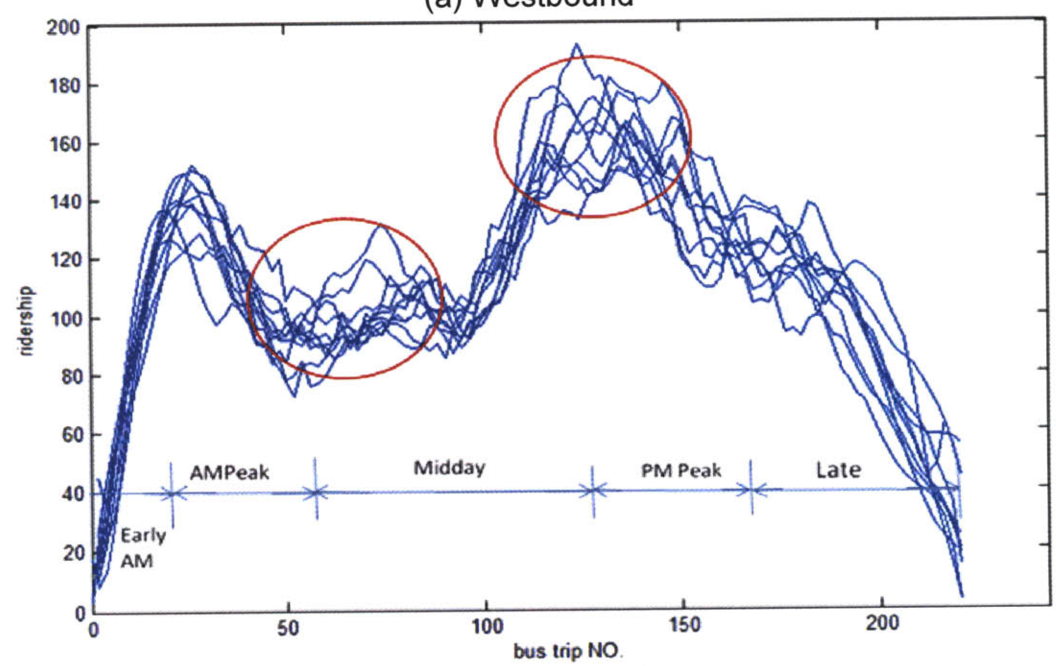
(b) Average of Bus Trips 127, 129, 131

Figure 5-7: Average trip-level ridership for Route 70 during PM Peak

Similar variation is found over five successive weekdays on Route 185 (May 11-15, 2009), as shown by the red circles in Figure 5-8. Similar variations were found for the other routes analyzed.



(a) Westbound



(b) Eastbound

Figure 5-8: Daily variation in trip-level ridership for Route 185

5.2 Daily Load Profiles

This section explores how the estimated passenger load differs between the expanded BODS counts and expanded Oyster counts. The BODS survey counts passengers for only one day every 5-7 years, whereas the proposed automatic procedure can produce OD matrices for any route on any day with adequate ADCS archived data. Therefore, based on the inference results, we can get more detailed information about the spatial and temporal variation in loads from the proposed procedure.

5.2.1 Expansion process for BODS

The BODS survey on any route usually takes place over the course of a single day. The number of bus trips sampled varies with route frequency, but roughly 50% of the total bus trips are surveyed. For the OD matrix estimation of non-surveyed bus trips, BODS looks at the next trip and, if that trip has been surveyed, it simply duplicates its number of total passengers and assumes that their OD patterns are also the same. If the next trip has not been surveyed, it looks at the previous trip and, if that has been surveyed, uses that. If it has not, it looks at the second following trip and, if that has not been surveyed, it then looks at the second previous trip, and so on. One shortcoming of this expansion method is that the estimation of the ridership for some bus trips that have not been surveyed may be influenced by bus bunching: if the ridership of the previous trip is higher than normal, the ridership of the following trip might be lower than normal.

5.2.2 Expansion process for Oyster data

For the Oyster process, around 10% of all bus passenger trips are non-Oyster trips (as shown in Section 5.1), and the origins and destinations for these trips cannot be inferred directly using the proposed method. Also, as discussed previously, not all Oyster trips can have origins and destinations inferred. So the Oyster "seed" OD matrix represents only about 60% of all bus passenger trips. Control totals are needed to expand the "seed" matrix to represent the total bus passenger trips. Since the ETM dataset records both Oyster and non-Oyster transactions, it can be used as the control total.

Given the Oyster "seed" OD matrix and the corresponding control totals, it is possible to scale up the "seed" matrix, provide an estimate of the actual ridership. The

Iterative Proportional Fitting (IPF) method has been used successfully in MIT's work estimating bus and rail stop-by-stop OD matrices for the Chicago Transit Authority network. Ideally, it would be used to scale the "seed" matrices up to match the control totals as closely as possible. However, in this case, the ETM dataset records only the number of boardings at the bus trip-level, rather than the stop-level. Therefore, a simpler method is used to scale up the OD matrix in this case.

Two assumptions are made to scale up the "seed" matrix:

1) Passengers on bus trip T who were inferred to have boarded at stop O (but for whom no destinations were inferred) are assumed to have the same destination distribution as passengers on bus trip T who also boarded at stop O but had destinations inferred. In this case, Equation 4-1 can be used to expand the "seed" matrix to all passengers:

$$F'_{ODT} = F_{ODT} + B_{OT} * \frac{\sum_T F_{ODT}}{\sum_T \sum_D F_{ODT}} \tag{4-1}$$

where B_{OT} denotes the number of passengers on bus trip T who were found to have boarded at stop O but for whom no destination stops could be inferred; and F_{ODT} denotes the number of passengers on bus trip T who were inferred to have origin O and destination D.

2) Non-Oyster passengers and Oyster passengers who have neither origin nor destination inferred are assumed to have the same origin and destination distribution as F_{ODT}'. The fully scaled O-D total is represented by V_{OD} in Equation 4-2:

$$V_{OD} = F'_{ODT} * \frac{T_T}{\sum_O \sum_D F'_{ODT}} \tag{4-2}$$

where T_T denotes the total number of boardings for bus trip T (based on the ETM dataset).The scaled OD total V_{OD} value is then further divided by direction and time period for the following load profiles and trip distance distribution comparisons with the expanded BODS OD flows.

5.2.3 Load profiles

Load profiles are standard graphics showing passenger activity (boardings, alightings) and passenger load at each bus stop along a route by direction. They allow planners to identify locations and values of the peak load, as well as any underutilized route segments. This section compares the load profiles derived from BODS and Oyster at the daily and time period level by direction for the selected routes.

1) Load profile comparison between BODS and Oyster

Load profiles are generated from the BODS and Oyster O-D matrices by direction and by time period. Routes 185 and 307 are chosen here as examples to show the comparison results. All the comparisons shown in this section focus on the days when the BODS surveys were conducted for the selected routes.

Since both the BODS and the Oyster results are expanded, the first step is to check the total ridership to make sure both the BODS and the Oyster estimates are expanded to the same control totals. However, significant differences were found between the total ridership by direction and time period. This finding is due to the different procedures used to expand the ridership to the control total: the Oyster estimates are expanded for each bus trip using the ETM data as control totals, while BODS counts are expanded by duplicating the closest adjacent surveyed bus trip to all non-surveyed trips (see Section 5.2.1). Tables 5-1 and 5-2 show the comparison results for Routes 185 and 307, respectively. For the other three routes, the ridership comparison results are included in Appendix C.

Table 5-1: Ridership comparison for Route 185

Timeband	Eastbound			Westbound			Total		
	BODS	ETM	ETM/ BODS(%)	BODS	ETM	ETM/ BODS(%)	BODS	ETM	ETM/ BODS(%)
Early AM	623	177	28%	1426	867	61%	2049	1044	51%
AM Peak	2180	1911	88%	2551	2315	91%	4731	4226	89%
Midday	4625	4744	103%	5316	5232	98%	9941	9976	100%
PM Peak	2621	2772	106%	2308	2762	119%	4929	5534	112%
Late	2887	3255	112%	2332	2424	104%	5219	5679	109%
All Day	12936	12859	99%	13933	13600	97%	26869	26459	98%

Table 5-2: Ridership comparison for Route 307

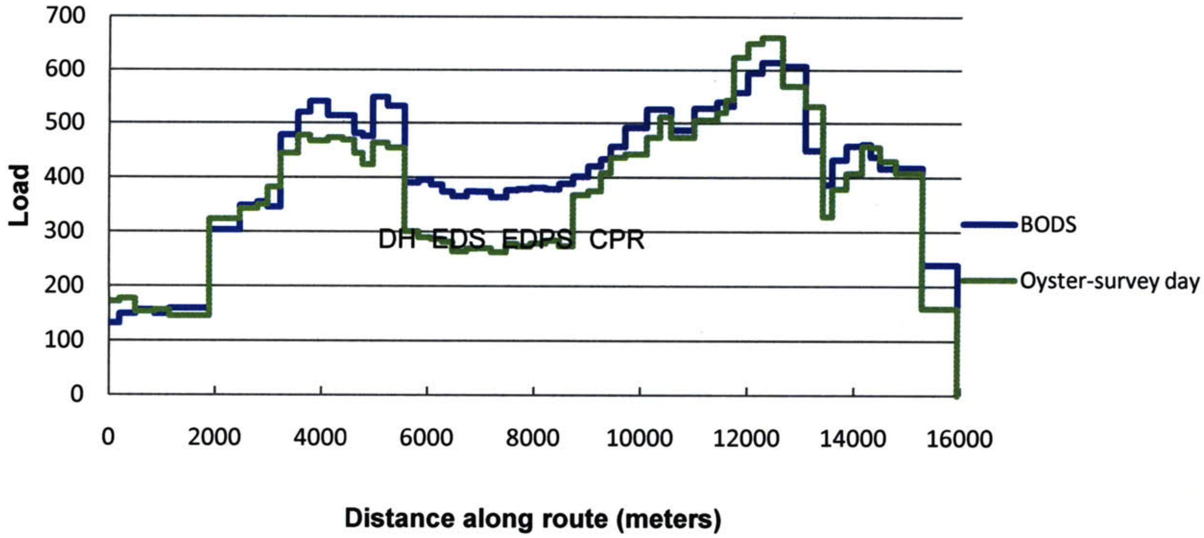
Timeband	Eastbound			Westbound			Total		
	BODS	ETM	ETM/ BODS(%)	BODS	ETM	ETM/ BODS(%)	BODS	ETM	ETM/ BODS(%)
Early AM	193	162	84%	569	261	46%	762	423	56%
AM Peak	1102	936	85%	1312	1231	94%	2414	2167	90%
Midday	2798	2288	82%	2834	2555	90%	5632	4843	86%
PM Peak	1058	1074	101%	823	938	114%	1881	2012	107%
Late	1118	956	86%	813	749	92%	1931	1705	88%
All Day	6269	5467	87%	6351	5683	89%	12620	11150	88%

The two datasets (BODS and Oyster) are expanded using different methods and as a result, some large differences in total ridership are found in some time periods. The Oyster-based ridership is mostly lower than the BODS figure, especially for Route 307. In particular, the Oyster-based ridership is far lower than the BODS figure during the "Early AM" period. One major reason for this is that during this time period (4:30 - 6:59 a.m.), the BODS survey typically starts at 6:25 a.m., and bus trips scheduled before 6:25 a.m. are assumed to have the same ridership per trip as those later surveyed trips. Clearly, it is unlikely that bus trips running before 6 a.m. will have the same ridership as the trips between 6:25 and 6:59 a.m. This explains why the expanded BODS data consistently overestimates ridership during the Early AM time period. Appendix B shows similar observations for the other three selected routes.

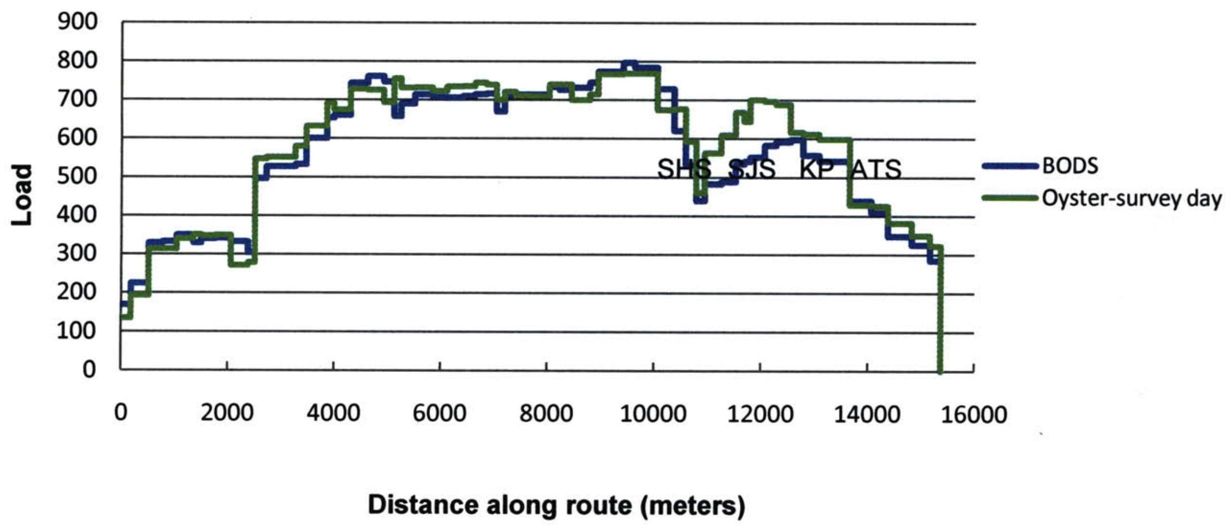
Though the total ridership of expanded BODS and Oyster do not match exactly due to different expansion methods as discussed above, load profiles are compared based on these estimates. Route 185 during the AM Peak period is chosen here to show the load comparison results (see Figure 5-9).

The key stations[2] where larger load differences occur between the expanded BODS and Oyster datasets are marked on the load profiles.

[2] DH=Denmark Hill (connection to National Rail) EDS=East Dulwich Station (connection to National Rail) EDPS=East Dulwich Police Station CPR=Crystal Palace Road SHS=Sacred Heart School SJS=St John's School KP=Kennington Park ATS=Archbishop Tenison's School



(a) Eastbound



(b) Westbound

Figure 5-9: Load profile for Route 185 in AM Peak

In general, the Oyster-based load profiles match the expanded BODS load profiles well, particularly eastbound. However there are significant differences between Denmark Hill bus station and Crystal Palace Road bus station (eastbound), and between Sacred Heart School bus station and Archbishop Tenison's School bus station (westbound). For Route 185 eastbound, the expanded Oyster loads in the AM Peak are 88% of the expanded BODS loads. There is not enough information to explain whether the expanded Oyster data underestimates or the BODS data overestimates the load on this segment. In the westbound direction, the main differences (where the expanded BODS loads are less than the Oyster loads) are concentrated around bus stops that are close to three schools. It is quite possible that the BODS survey does not capture all the passengers at these bus stops, especially in the AM Peak, when many students are traveling.

2) Load profile variation across days

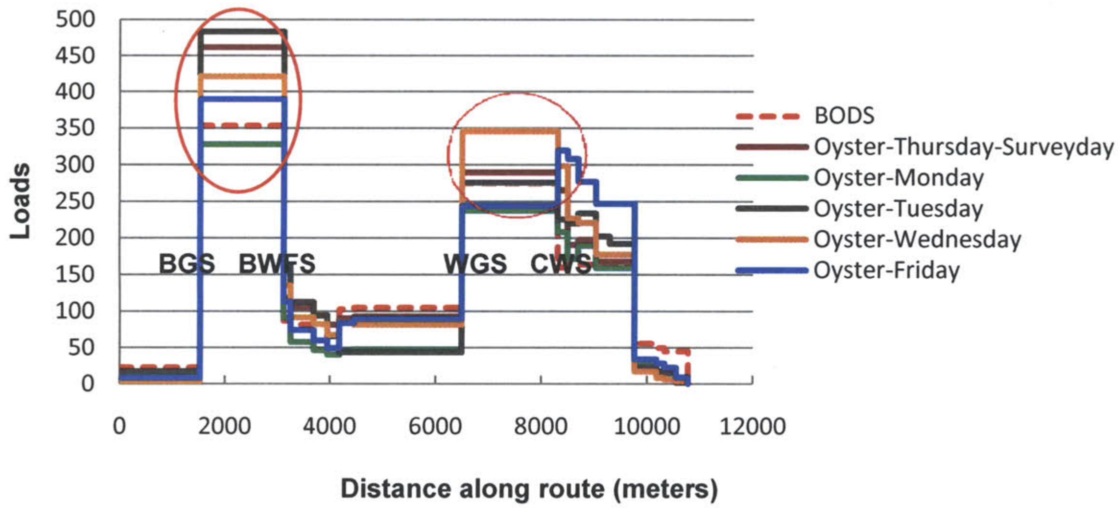
Route W4 during the AM Peak is chosen here as an example of how the daily load profile varies over five successive weekdays. Figure 5-10 shows that there are large variations in the load profile and specifically in the peak loads, even within the same week.

On the BODS survey day, the Oyster load profiles are generally consistent with BODS, except the Oyster peak loads are much higher than BODS, specifically in the northbound direction, where the expanded Oyster control total is 106% of the expanded BODS control total. It is likely that the BODS survey underestimates the peak loads due to the low sample rates at peak periods or the effect of the BODS expansion method.

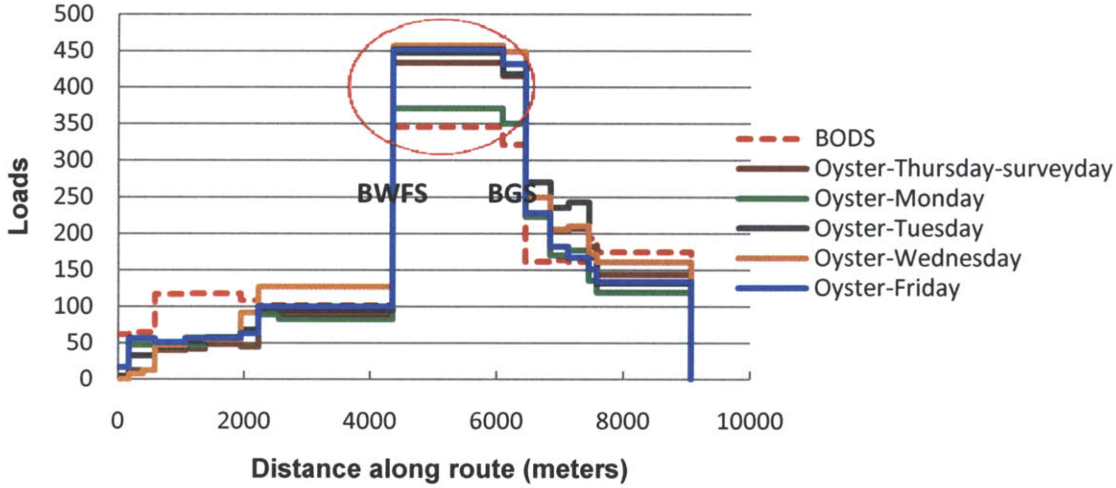
It is also interesting to note that in both directions along Route W4, Monday seems to have the lowest maximum loads. The locations for the peak points are consistent for both directions across days, mainly between the Bruce Grove Station and the Broadwater Farm Estate bus stop.[3]

In general, the load profiles vary across days. Though the Oyster estimates are consistent with BODS on the survey day, the load profiles from Oyster on the other weekdays are much different from BODS. Therefore, it appears that the one-day survey is not enough to provide reliable information about the daily load profiles.

[3] CWS= Chequers Way Station
BWFS= Broadwater Farm Estate
WGS= Wood Green Station (close to a shopping center)
BGS=Bruce Grove Station (connection to National Rail)



(a) Northbound

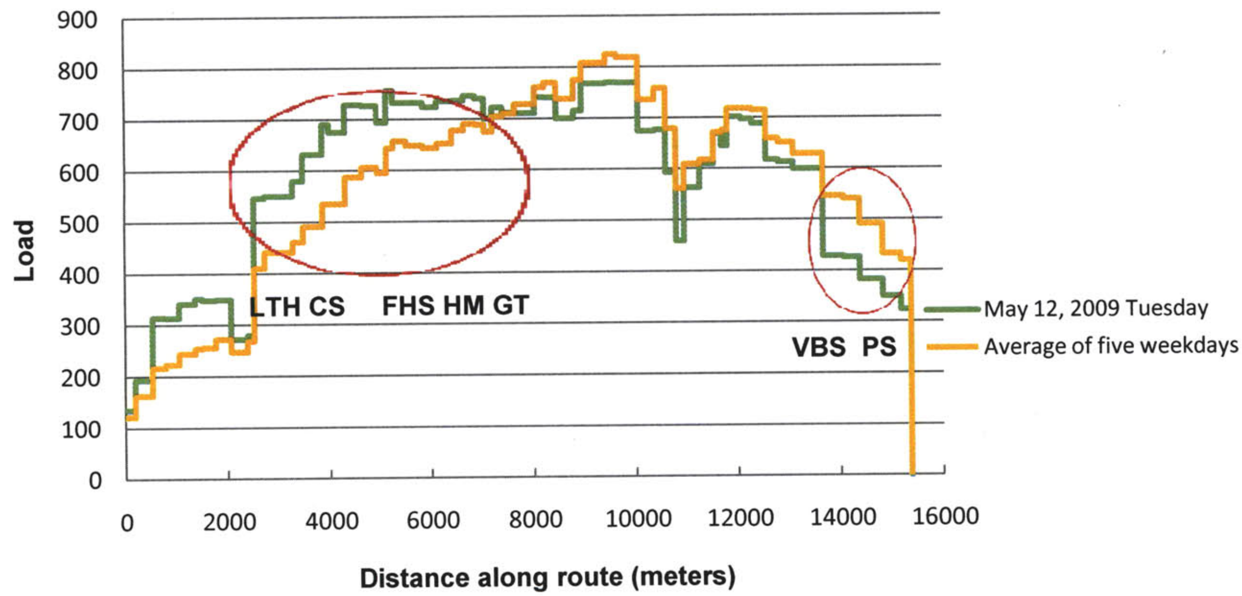


(b) Southbound

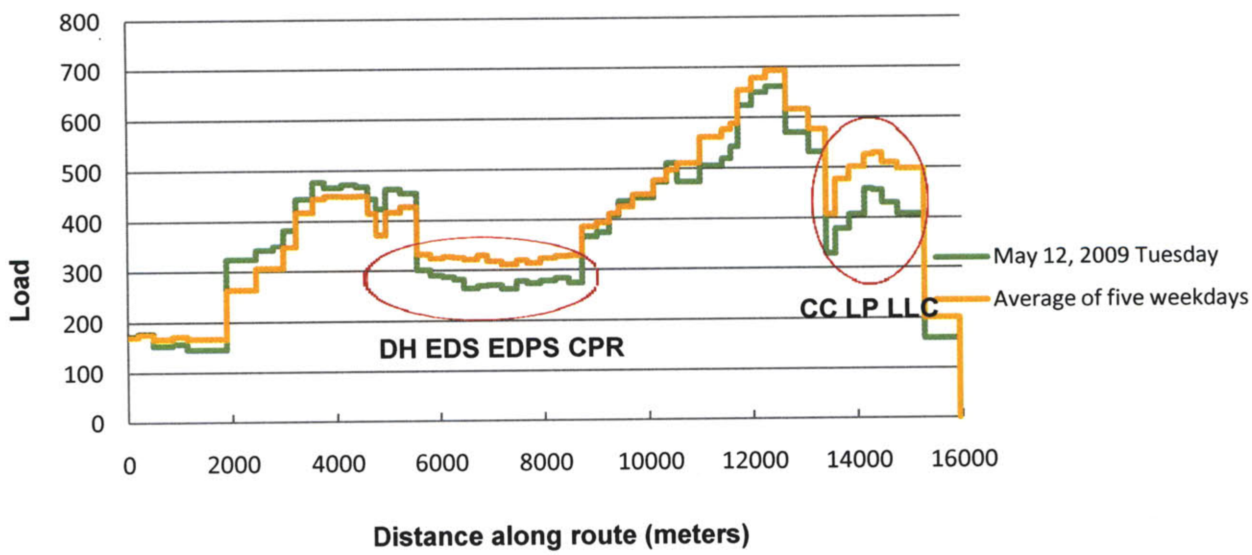
Figure 5-10: Daily load variation along Route W4 during AM Peak

3) Load variation between a single weekday and the average weekday

Route 185 is chosen here to show the load variation between a single weekday (the BODS survey day, i.e., May 12, 2009) and the average of five successive weekdays. Figure 5-11 again shows that a one-day load sample cannot reliably represent the typical weekday: large load variations can exist between a single weekday and the average of five successive weekdays. These variations are often greatest at bus stops with National Rail connections, in the Central Business District or at entertainment sites[4].



(a) Westbound



(b) Eastbound

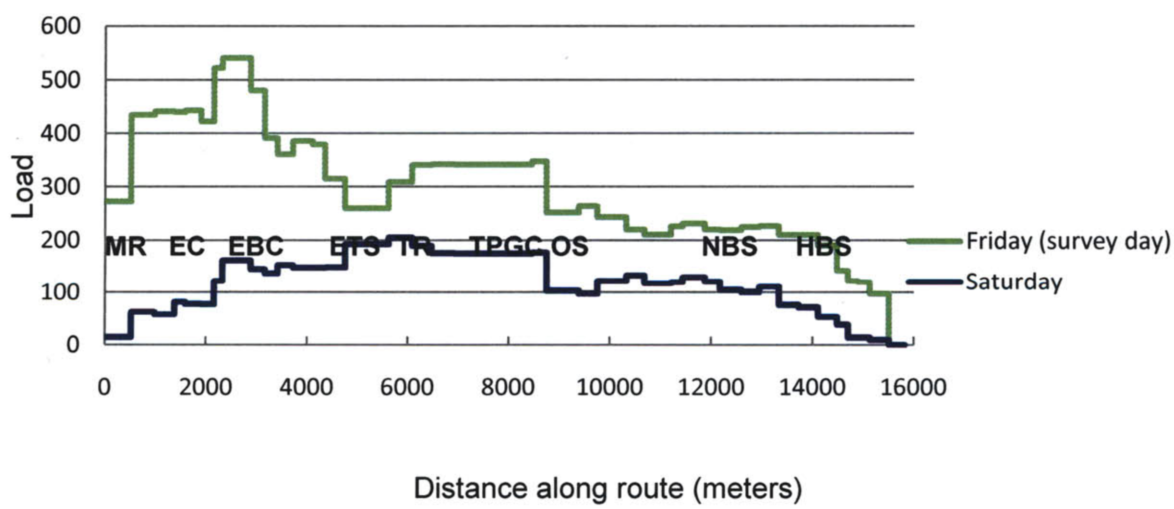
Figure 5-11: Load profiles on Route 185 in AM Peak

[4] LTH=Lewisham Town Hall CS=Catford Station (connection to National Rail)
FHS=Forest Hill Station(connection to National Rail) HM=Horniman Museum GT=Grove Tavern
VBS=Victoria Bus Station PS=Pimlico Station DH=Denmark Hill (connection to National Rail)
EDS=East Dulwich Station (connection to National Rail) CPR=Crystal Palace Road
CC= the Catford Center LP=Lewisham Park LLC=Ladywell Leisure Center

4) Load variation between weekdays and weekends

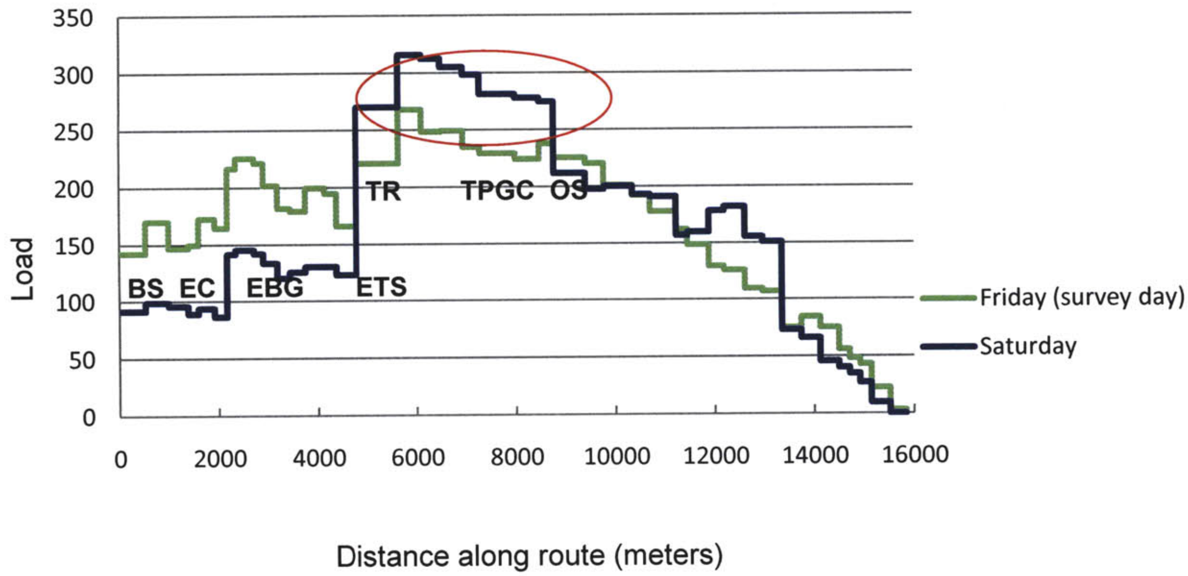
Since the OD information can be obtained for every day with the ADCS archived data, we can also study the load profile differences between weekdays and weekends. Figure 5-12 demonstrates the load profile variations between a Friday and Saturday on Route 307. Even though these are successive days, the loads are different on this route. Generally, the load on Saturday is much lower than that on Friday, and the peak load point changes in the AM Peak. On Friday, the peak load point is between Glyn Road[5] and Crown Road while on Saturday, the peak load point is between Enfield Town Station, Trent Park Golf Course and Oakwood Station. It is quite likely that more passengers visit shopping and recreation destinations on Saturday, which changes the peak load segment of this route.

In the PM Peak, as shown by the red circle in Figure 5-12 (b), the peak load points are also around the Trent Park Golf Course bus stop and Oakwood Station, and the loads around these stops are even larger than on Friday. It is likely that more people may transfer at Oakwood Station to other routes or the Underground at weekends for non-work trip purposes.



(a) AM Peak (Westbound)

[5] BS= Brimsdown Station MR=Mayfield Road EC=Enfield College EBG=Enfield Bus Garage ETS=Enfield Town Station TR=The Ridgeway TRGC=Trent Park Golf Course OS=Oakwood Station NBS=New Barnet Sainsburys HBS=High Barnet Station



(b) PM Peak (Westbound)

Figure 5-12: Load profiles for Route 307

Similar results for Route 329 (attached in Appendix B) support the same conclusion that the load profile varies greatly between weekdays and weekends.

Since the daily loads vary greatly and one day's information may not be representative of other days, it is very useful if transportation planners can get such information daily or on an average day basis. With the ADCS archived data and the inferred OD matrices, it is straightforward to get these daily load profiles which are infeasible with traditional surveys. This opens the door to improving service planning by basing it on more extensive and reliable information.

5.3 Passenger Trip Distances

Average passenger trip distances are compared between BODS and Oyster estimates as another check on the consistency of data across these two data sources. Route 70 is chosen here with the comparison results shown on the BODS survey day in Table 5-3.

Table 5-3: Average distance traveled along Route 70 (kilometers)

Time Period	Eastbound			Westbound		
	BODS	Oyster	Oyster/ BODS (%)	BODS	Oyster	Oyster/ BODS (%)
Early AM	3.8	4.1	108%	2.5	2.0	80%
AM Peak	2.7	2.9	107%	2.5	2.4	96%
Midday	2.8	2.7	96%	2.6	2.5	96%
PM Peak	2.8	2.6	93%	2.8	2.5	89%
Late	2.9	3.0	103%	3.1	3.0	97%
All Day	2.9	2.8	97%	2.7	2.5	93%

Comparison of the average distance traveled between BODS and Oyster estimates shows fairly consistent results with the following exceptions:

- Early AM westbound, where the Oyster distance is 20% shorter than BODS, which may be attributed to the variance associated with a small BODS sample size and the definition of the time period;
- PM Peak westbound, where the Oyster distance is 11% shorter than BODS; it is hard to judge which data source is not accurate due to a lack of information.

For the other four selected routes, the comparison results (see Appendix B) support the conclusion that the average distances traveled are generally consistent between the expanded Oyster estimates and the expanded BODS data.

5.4 Interchange Time Analysis

Understanding the behavior of public transportation customers may allow transit agencies to provide riders with a better experience. Interchanges affect the attractiveness of public transportation and making interchanges less burdensome is a critical consideration in public transport planning. Improving the level of service for interchange locations would enhance the overall quality of public transportation services. If the stops or stations that have heavy interchange demands but long interchange times can be identified, then improvements may be possible to enhance passengers' convenience.

Both practitioners and researchers tend to pay most attention to the initial waiting experience and to in-vehicle travel for their obvious effects on ridership, but less work has been done on interchanges between segments of a linked journey. However, reducing the out-of-vehicle times can help make public transit more attractive, resulting in ridership increases. Also, passengers may stop using transit service because of poor connections. A common rule of thumb is that walking and waiting time are considered by transit users to be two to three times as onerous as in-vehicle travel time. Designing routes and schedules with a minimum amount of waiting time during interchanges may decrease the level of inconvenience.

Theoretically, transport service providers can see when a smart card has been used to interchange between two (or more) buses within a defined time period and identify where the interchange took place. The service provider can then, for example, make a service planning decision on whether a through bus service should be introduced to cater to the groups of users interchanging between certain bus routes. This section will review the previous studies related to interchange times and then propose a new method to analyze bus passengers' interchange times using London as an example.

5.4.1 Prior Work

Hofmann and Mahony (2005) develop an iterative algorithm to classify passenger boardings into interchange journeys and single journeys. The electronic fare collection dataset used is from an urban public transport operator with a large fleet (over 1000 buses) and data for 48 million magnetic stripe card boardings from 1998 and 1999. This research describes the automatic generation of a new data attribute that cannot be derived directly and therefore increases the future utilization of the dataset. They restrict interchange journeys to those with a time difference between two successive bus boardings of no more than 90 minutes (note that this includes both the in-vehicle travel time on the first bus trip as well as the interchange wait time for the second bus). In their analysis, the time difference of the middle 50% (between 25 percentile and 75 percentile) of all interchange journeys are between 20 and 53 minutes. The median passenger interchange time is found to be 34 minutes. And 10% of all interchange passengers spend between 72 and 89 minutes after boarding the first bus, as shown in Figure 5-13.

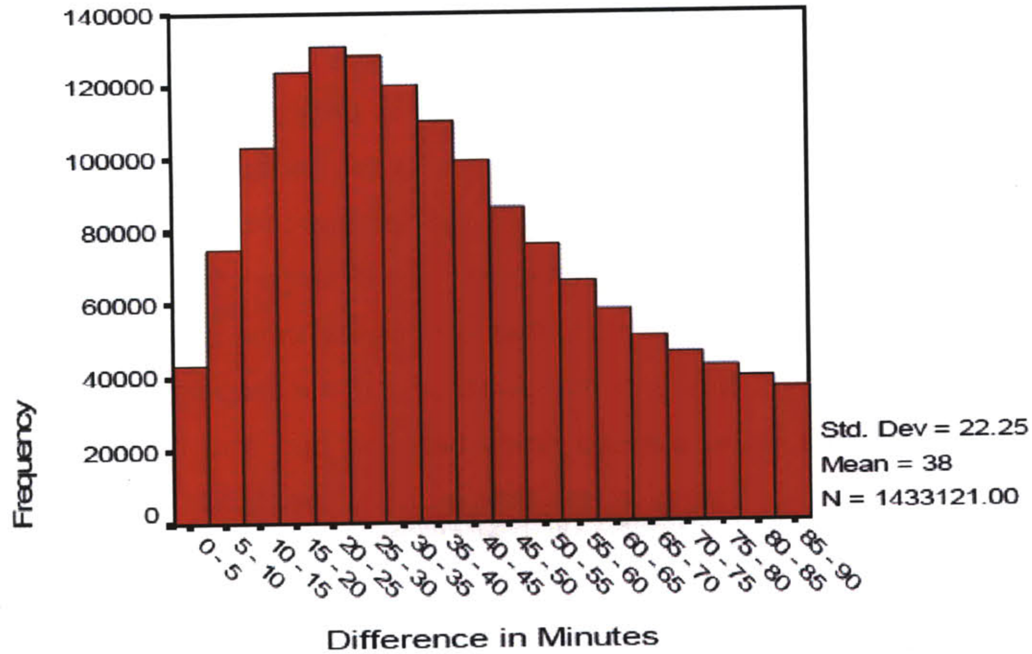


Figure 5-13: Histogram of successive boarding times difference (Hofmann & Mahony, 2005)

Chu and Chapleau (2008) used Trépanier et al. (2007)'s proposed alighting stop inference method to infer alighting stops for bus passenger trips. Based on that, they further use the additional knowledge of scheduled departure and arrival times for each bus run, the stop sequence, and the linear distance between stops to estimate the alighting time and identify linked trips by using spatial-temporal concepts. In their study of National Capital Region of Canada, with most routes connecting Gatineau, Québec, and central business district of Ottawa, Ontario, they found the median bus-to-bus interchange time was 7 minutes, and more than 80% of linked trips had interchange times of 18 minutes or less.

Seaborn (2008) summarized previous approaches for interchange time analysis. For example, Bagchi and White (2004) link two bus passenger trips that begin within 30 minutes of each other as recorded by smart card transactions in their study area around Bradford in West Yorkshire, UK, but assert that in larger cities a wider time window would be needed to identify complete trips. Okamura, Zhang and Akimasa (2004, cited in Seaborn 2008) define an interchange as two journey stages that are provided by two different operators within a waiting time of 60 min at the same stop. They use this definition to analyze interchange waiting times at major transit hubs.

Another application of the ADCS archived data to interchange time analysis is provided by Seaborn (2008), who examines three potential interchange combinations to gain an understanding of interchange behavior in London and recommends elapsed time thresholds for identifying interchanges across the London Network : 20 minutes for Underground-to-bus interchanges, 35 minutes for bus-to-Underground interchanges, and 45 minutes for bus-to-bus interchanges with a range of values that account for variability across the network. She further summaries three additional findings about interchange behavior for bus and the Underground passengers in London. She first points out the variations in elapsed times between bus and the Underground, and between two bus journey stages across time periods, specifically in the Midday and PM Peak periods. Secondly, she finds that variations of elapsed times between journey stages at different Underground stations appear to be closely related to the land use patterns around stations. She also finds that bus passenger interchange behavior is influenced by ticket type, for example Pay-As-You-Go (PAYG) versus pass holders. In her research, the complete passenger journeys are identified using the elapsed time thresholds, which are estimated from smart card data. No additional data from the iBus AVL systems is used in this research.

Though her research provides new and relevant interchange information that supports network planners with a more integrated view of bus service performance, the spatial accuracy of bus boardings is limited to the bus route-level, and no information about alightings can be obtained. Therefore, the interchange time in her research is actually journey time plus the “true” interchange time, so the in-vehicle journey time and walking/waiting time cannot be distinguished.

Jang (2010) presents a process to generate a travel time map for public transit system using smart card data. In his study, the AFC system in Seoul records each trip’s boarding and alighting times and locations as well as the trip chains with interchanges. Hence, the spatial and temporal distribution of interchanges can be obtained easily, which helps to distinguish the locations where interchanges occur from other stops or stations. He finds that more than 80% of interchange trips have an interchange time of 10 min or less in Seoul.

5.4.2 Methodology

Previous attempts have been made to analyze linked trips using the AFC data. Apart from determining if the routes taken are different, the identification of linked trips in previous studies is solely based on a fixed temporal threshold between two successive transactions. The weakness of applying a threshold value is that it can be seen as arbitrary even with consideration of network size and specific user groups. A fixed value that does not take into account the in-vehicle travel time and route headway would invariably classify all the boardings that are carried out within the threshold as linked trips. Trips with a short duration or trip chains might be masked and return trips might also be counted as interchanges.

In this research as discussed earlier, bus passengers' alighting locations can be inferred from the ADCS archived data. Also, since the iBus AVL data provide information about the observed departure time for each bus at each stop, by matching the inferred alighting locations with the iBus AVL data, we can estimate the alighting time for the individual bus passenger trip. Hence, the interchange time can be calculated more accurately as the difference between the subsequent trip's boarding time and the previous trip's alighting time.

5.4.3 Case Study of Interchange Time Analysis in London

London has more than 600 stations where passengers can change between different transportation modes. TfL listed the following themes to improve the interchanges:

(1)Efficiency: Operations, moving around, sustainability;

(2)Usability: Accessibility, safety and preventing accidents, personal security, protection;

(3)Understanding: Legibility, permeability, wayfinding, information;

(4)Quality: Perception, design, spaces, sense of place." (Transport for London)

Efficiency is one of the most important attributes of any interchange facility to attract riders, but there are no detailed statistics describing the current London transit passengers' interchange experiences. In this section, two potential interchange combinations (bus-to-bus and bus-to-Underground) are examined to gain an understanding of interchange behavior for London bus passengers and to formulate recommendations for "actual" interchange time to evaluate the current level of service for buses in the TfL network. "Actual" interchange times for passengers who take Routes 185, 307 and 329 are used as examples to illustrate bus-to-bus and bus-to-Underground interchanges.

1) Bus-to-Bus Interchange

Bus-to-bus interchanges can be divided into two cases: one is to interchange from one bus route to a different bus route, and the other is to "interchange" within the same bus route. The latter includes mostly return trips. The "interchange" times for the second case (i.e., return trips) are often much longer than the first case. Figure 5-14 shows the cumulative distributions for return trips on Route 185 on the BODS survey day. The time differences (i.e., the next trip's boarding time – the previous trip's alighting time) for such return trips are generally more than 2 hours.

These return trips are excluded from the bus-to-bus interchange time analysis since the large time differences indicate that these trips are unlinked trips with different travel purposes, and thus, they are not interchanges. For the first case of interchanging from one bus route to another, since the "actual" interchange time is the difference between the subsequent trip's boarding time and the previous trip's alighting time, not only does the alighting time from the previous trip matter, but also the headway of the connecting bus route. For example, it is quite possible that when a bus passenger gets off a bus along Route 185, he/she may just miss the bus on the connecting route and have to wait a full headway for the next bus. Or a passenger, especially an elderly one finds the first arriving bus to be crowded with no seats available, and decides to wait for the next bus to get a seat. However, if the difference between the previous trip's alighting time and the subsequent trip's boarding time is much larger than the headway of the connecting bus route, it is quite likely that these two successive trips are not linked trips. It may be this passenger has some other activities around the alighting stop, and does not want to take the connecting bus immediately. These trips are excluded from this interchange study. Therefore, when we judge whether two successive trips are linked trips or not, we consider not only the cumulative interchange time distributions, but also the headways of the connecting routes.

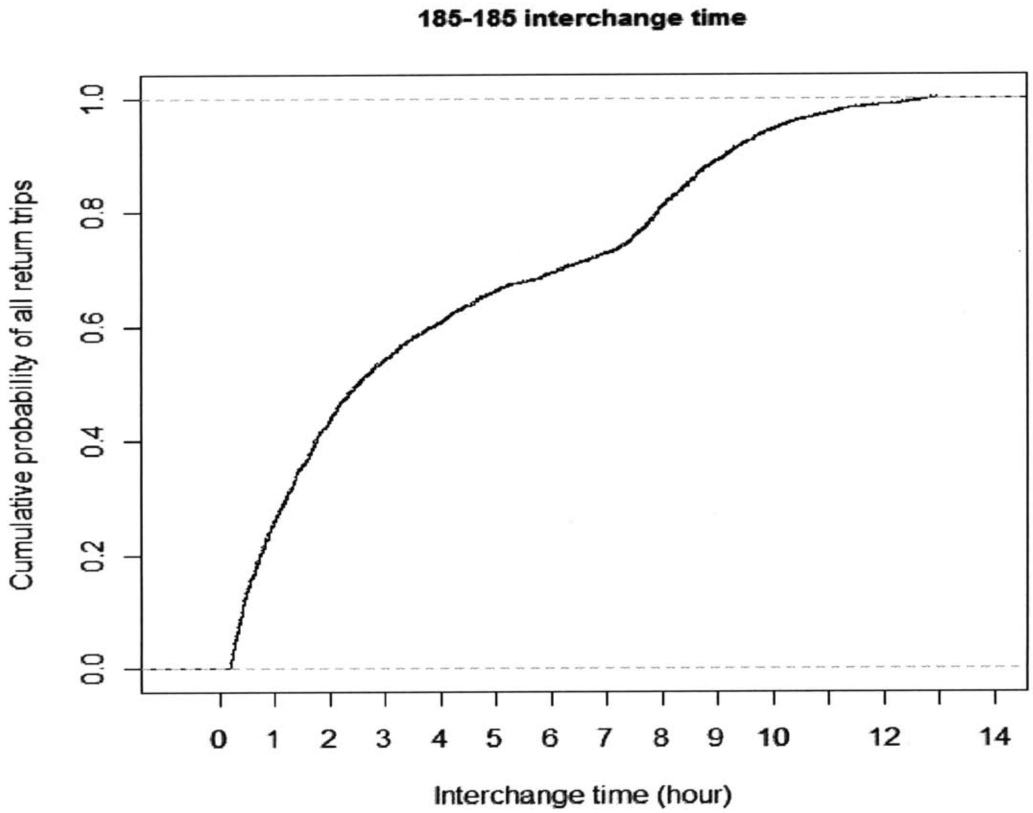


Figure 5-14: CDF of interchange times for return trips on Route 185

For example, 12.4% of the passengers on Route 185 reboard on Route 185 for the following transit trip (mostly return trips as discussed previously), while 50.8% of them take other bus routes. Figure 5-15 shows the cumulative interchange time distribution for these 50.8% passenger trips on Route 185. It is difficult to find the threshold value for bus-to-bus interchanges based on this graph. The headway distributions for these passengers' subsequent bus trips are further checked to help define the threshold value. We can see from Figure 5-16 that most of the headways for the connecting bus routes are within 15 minutes.

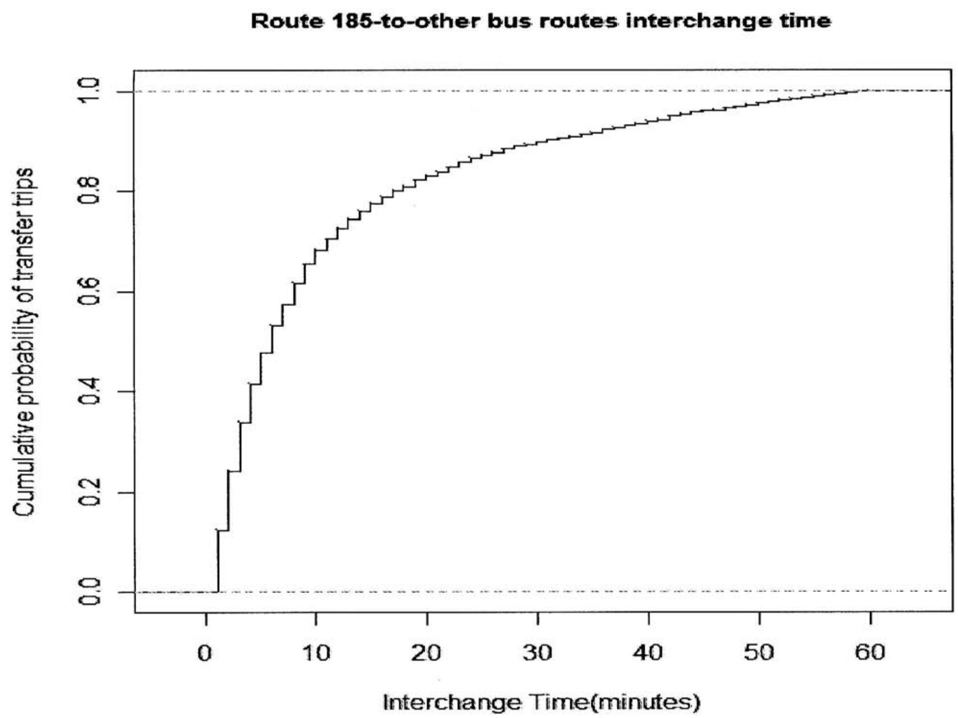


Figure 5-15: CDF of bus-to-bus interchange times for passengers on Route 185

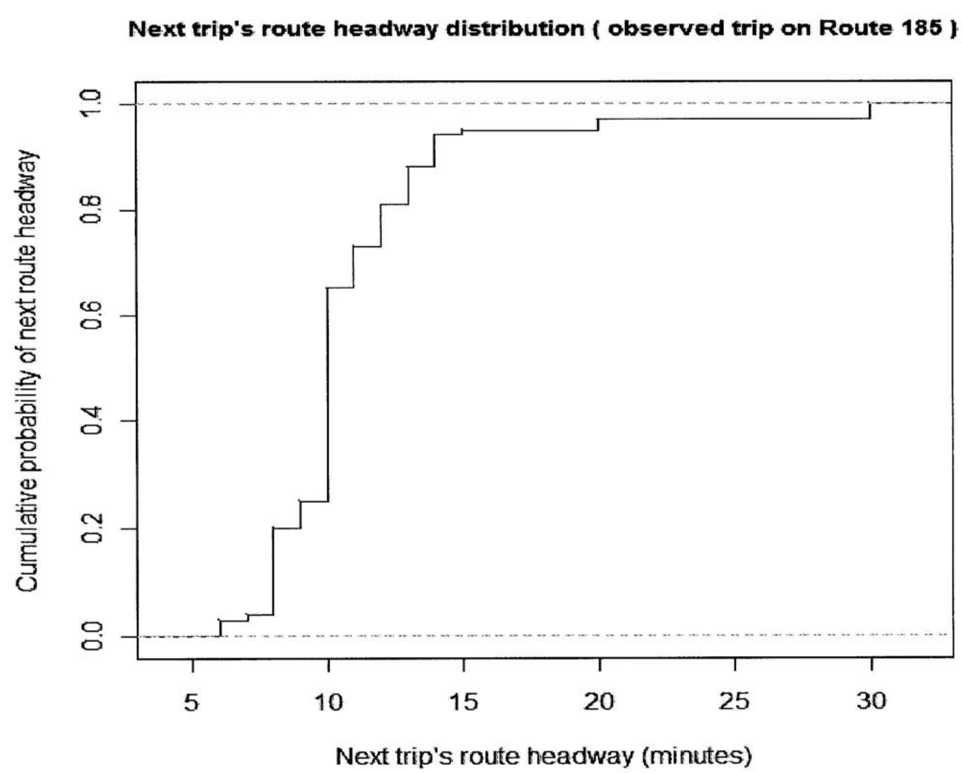


Figure 5-16: Connecting route headway distribution for passengers from Route 185

Therefore, 15 minutes is used here as the threshold value to distinguish between linked and unlinked journeys. For all trips originating from Route 185, if the time difference between alighting from Route 185 and the next boarding on another bus route is within 15 minutes, then these two trips are considered linked trips. For all such linked trips that meet this time threshold, the median bus-to-bus interchange time for passengers originating from Route 185 is about 4 minutes.

This method is based on the statistical analysis of the connecting bus route's headway distribution for all the passengers originating from the selected routes. A simplified method is proposed here and the results of the two methods are compared to check the feasibility of this simplified method. Recall that when we infer the origin, we need to check the GIS files to list all the routes that run parallel to or intersect with the studied route. Therefore, we know all the related routes that passengers on the studied routes may change to for their subsequent trips. The simplified method is based on this information and picks the maximum headway from all of these routes as the threshold to judge whether two trips are linked or unlinked (but it excludes those connecting routes that operate only at night). This value for Route 185 is 20 minutes while it is 15 minutes by doing the detailed analysis as mentioned above. The median interchange time for passengers originating from Route 185 using this simplified method is 5 minutes while it is 4 minutes by the detailed statistical analysis. Additional comparison results are shown in Table 5-4. The threshold values obtained by these two methods are similar, and thus it is reasonable to use this simplified method for planning purposes.

Table 5-4: Median interchange time

Current Route	Threshold by detailed analysis	Median interchange time by detailed analysis	Simplified threshold	Median interchange time using simplified threshold
185	15 minutes	4 minutes	20 minutes	5 minutes
307	20 minutes	6 minutes	30 minutes	7 minutes
329	15 minutes	5 minutes	20 minutes	6 minutes

More detailed analyses for the route-to-route and route-to-stop combinations are provided here as examples to demonstrate the micro level of the interchange time patterns. Taking Route 185 as an example, based on the Oyster transactions, Route 176 is found to be the most frequently used connecting route for passengers originating from Route 185, with 15 interchange stops for the parallel segments. The median interchange time for passengers from Route 185 to Route 176 is 5 minutes. The most frequently used connecting stop on Route 176 is the Forest Hill Station (which is a stop shared by both of these two routes, so interchange times for interchanges at this stop do not include walking time), with 7 minutes as the median interchange time for passengers originating from Route 185.

Transit planners often use half the headway of the connecting bus route as the estimated waiting time, but there is no field data to support this practice. The analysis in this part supports this assumption that the actual waiting time is approximately half the headway of the connecting route, as shown in Figure 5-17. The size of these dots indicates the number of interchange passengers on those connecting routes. Besides, by comparing the median interchange time with the headway of the connecting route, we can further evaluate the connecting services provided by those connecting bus routes. To be more specific, the dots below the diagonal line indicate the connecting bus routes with good connecting services while the dots above the diagonal line indicate bus routes with poor connecting services. In this case, for example, Routes 36 and 436 provided good connecting bus services for passengers originating from Route 185, while Routes P13 and 356 provided poorer connecting services. Thus targeted improvements could be made to coordinate the timetable. For this example on Route 185, the connecting services are fairly good as the median interchange times are approximately half the headway of the connecting routes.

Robustness is one of the major reasons why the median but not the average of the interchange times is chosen here to represent the typical bus-to-bus interchange patterns. Since the maximum headway among all the related bus routes is chosen to be the threshold to judge linked and unlinked trips, there is a risk that the route that has the largest headway may turn out to have many fewer transfer riders originating from the route under study. So choosing the headway of such a route as the threshold may not be representative of most of the linked trips originating from the studied route. However, the median value of the interchange times does not vary greatly, due to the change of the chosen maximum headway, even if the route with the maximum headway has many fewer riders. Thus it offers a robust way to demonstrate the interchange behavior.

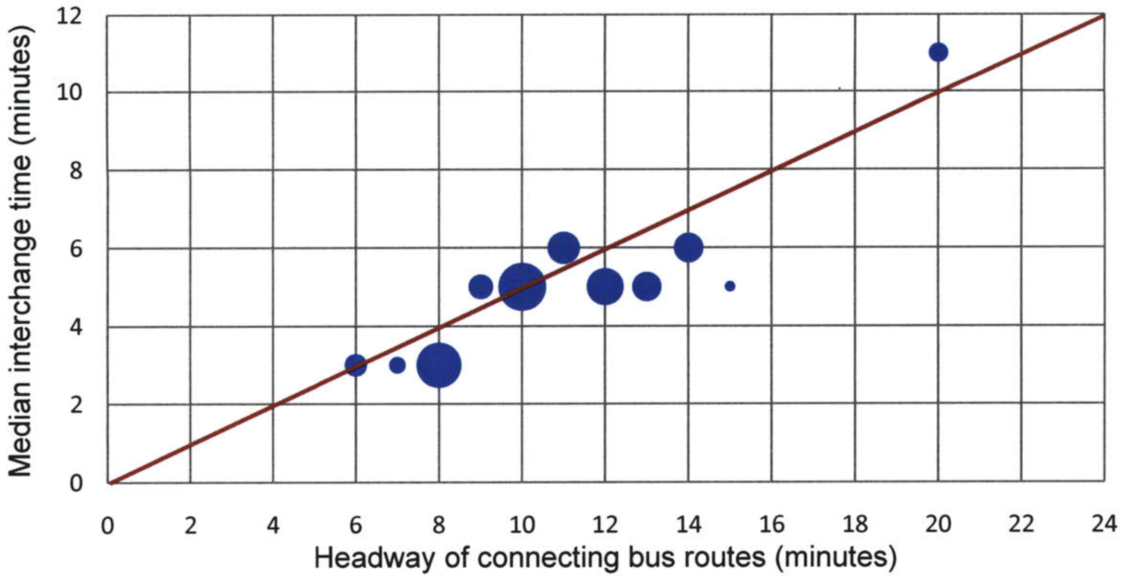


Figure 5-17: Relationship between connecting routes' headway and median interchange times

Moreover, from the perspective of transit agencies, the median interchange time for journeys originating from the selected routes can provide a rough assessment of the intermodal and intramodal connectivity for these routes (also an evaluation of the level of service). Based on this median value, transit planners can decide whether it is necessary to do more detailed interchange time analysis for the route-stop or route-route combinations to prioritize timetable coordination. Ideally, we can also study the various attributes of all bus stops (like the interchange demand and interchange time at each bus stop) and interchange facilities (for example, whether or not there are commercial uses of the facilities), to get a comprehensive user's perspective on the walking, waiting and interchange experiences. Therefore, transit agencies can provide passengers with better connecting services by shifting the bus schedules, relocating bus stops or providing through routing. However, recommendations of this type are beyond the scope of this research. This research calculated the median bus-to-bus interchange times to provide a rough assessment of the connection services that the selected routes can provide. The case studies selected in this thesis are intended to be illustrative rather than to provide a comprehensive assessment of bus passenger interchanges in London.

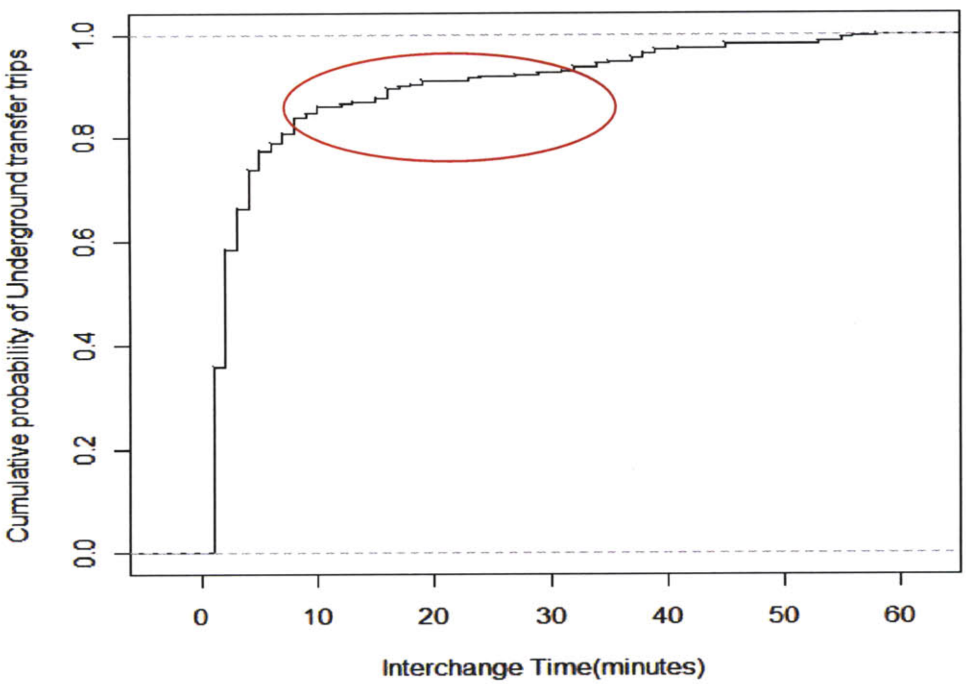
TfL provides a journey planning tool on the website for customers to get an idea of the estimated journey time between any OD pairs. The journey planning tool also includes the approximate walking and waiting time from one bus stop to another bus stop. The median bus-to-bus interchange times calculated by the method proposed here could provide additional customer information to check the reliability of the current estimated walking and waiting time posted on the TfL website.

2) Bus-to-Underground Interchange

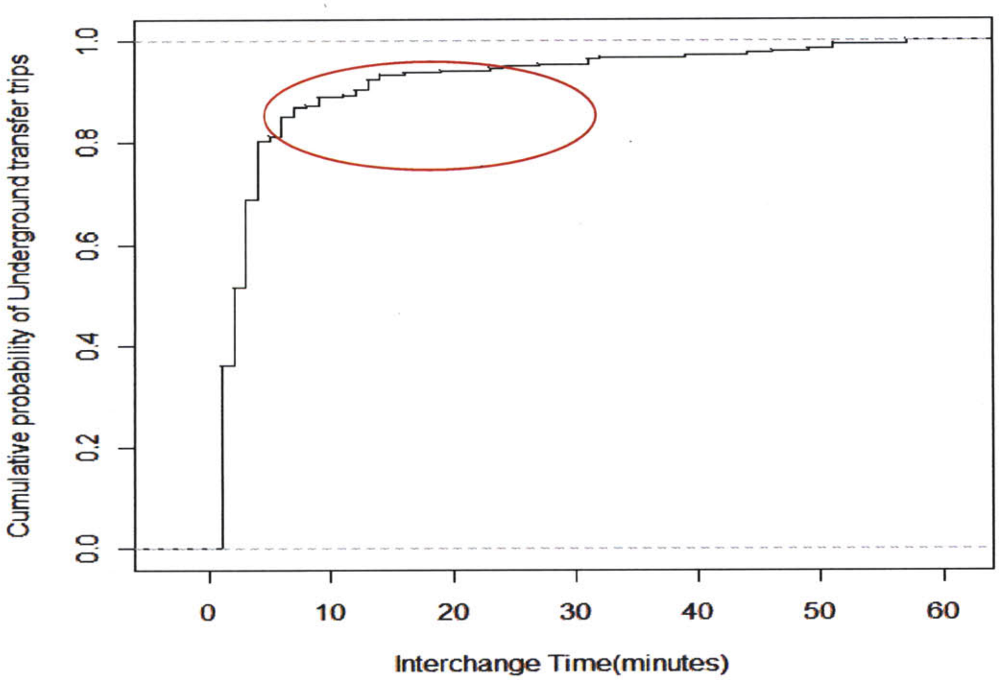
The bus-to-Underground interchange is different from bus-to-bus interchange, as bus-to-bus interchange time includes the time passengers walk to the next bus stop (or stay at the same stop where passengers get off from the previous trip) and the time that passengers wait at the bus stop platform for the connecting bus, which makes the interchange time dependent on the headway of the connecting bus route. However, for the bus-to-Underground interchange time, it is the difference between the subsequent trip's card tap time in the Underground or rail station and the alighting time for passengers from the previous bus, which excludes the platform waiting time.

Figure 5-18 shows the cumulative distribution function (CDF) for bus-to-Underground interchange times for passengers originating from Routes 185, 307 and 329. It shows that the CDF curves for bus-to-Underground interchange times are steeper than for bus-to-bus interchanges. The explanation is that most bus interchange stops do not have any commercial attractions and passengers stop here just to board the connecting buses. While most of the connecting Underground stations have shops or food courts, where passengers may stop for a while, which make the connecting Underground trip not a pure interchange trip. For passengers originating from Route 185, the bus-to-Underground CDF curve breaks at 6 minutes, with the median interchange time for interchanges meeting this threshold value being 2 minutes. Since platform waiting time is not counted in the bus-to-Underground interchanges, it makes sense that the median interchange time is shorter than the bus-to-bus case. The median bus-to-Underground interchange time for trips originating from other selected routes are all around 2 to 3 minutes.

(a) Route 185-to-Underground Interchange Time



(b) Route 307-to-Underground Interchange Time



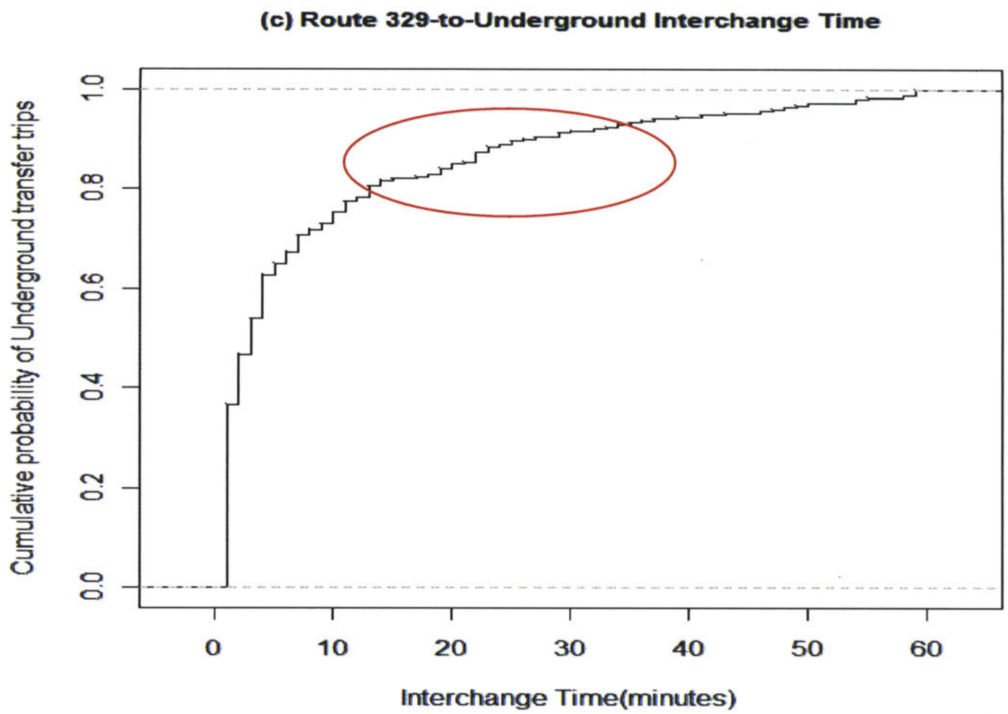


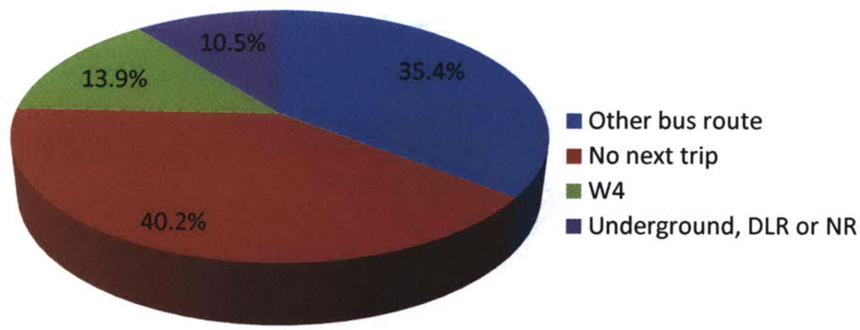
Figure 5-18: CDF of bus-to-Underground interchange times

The smoother parts of these bus-to-Underground interchange times curves shown by the red circles in Figure 5-18 can be understood by the land use patterns around some Underground stations. For example, many Underground stations have shops either within the station or nearby, and people may stop to buy a cup of coffee or go shopping; some smaller Underground stations have nothing close by, so people just interchange there without any other activities.

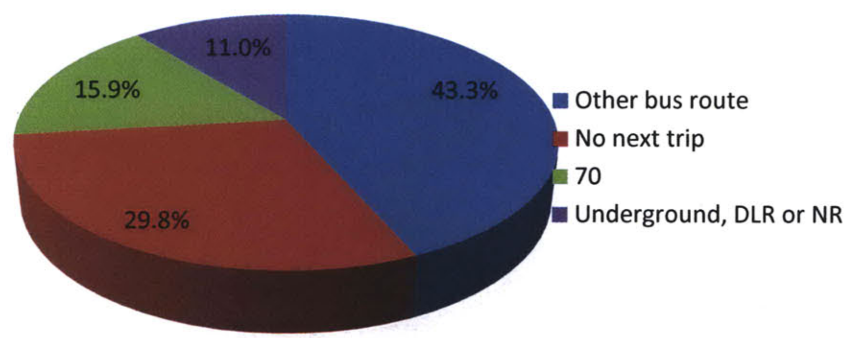
When locating bus stops, planners have generally considered making the bus stops and Underground stations as close as possible, so there is not necessarily much more improvement that can be made to the bus-to-Underground interchange stop locations. However, TfL can incorporate this information into their customer information tools to provide passengers a sense of the time that previous passengers took to interchange from bus to the Underground at specific stations.

5.5 Subsequent Travel Behavior for Bus Passengers

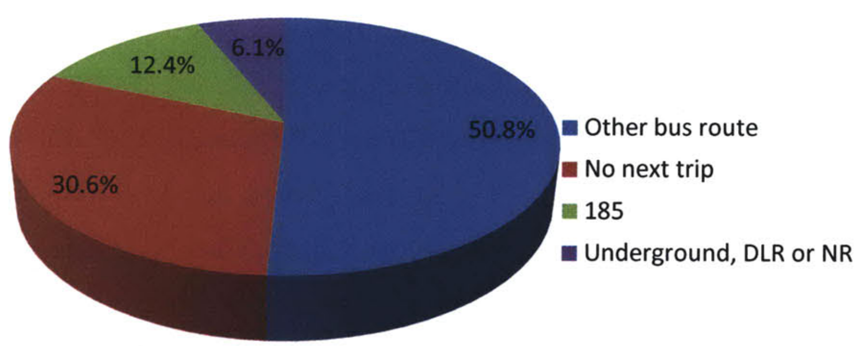
The previous section discussed the characteristics of linked trips, especially the interchange patterns. This section extends the discussion to all trips, including both linked and unlinked trips. Since the destination inference is based on the boarding information of the subsequent transit trip, this section provides a picture of the travel modes used for the next transit trips by Oyster passengers whose observed trips are on the selected bus routes. Figure 5-19 demonstrates the travel mode split of the subsequent trips for passengers riding on the selected routes on the BODS survey days. Note that all successive trips are analyzed and there is no restriction on the time difference between two boardings; namely, these analyzed successive trips are not necessarily linked trips.



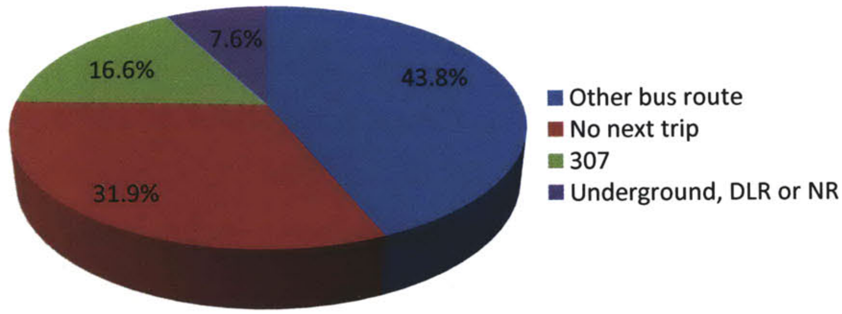
(a) Route W4



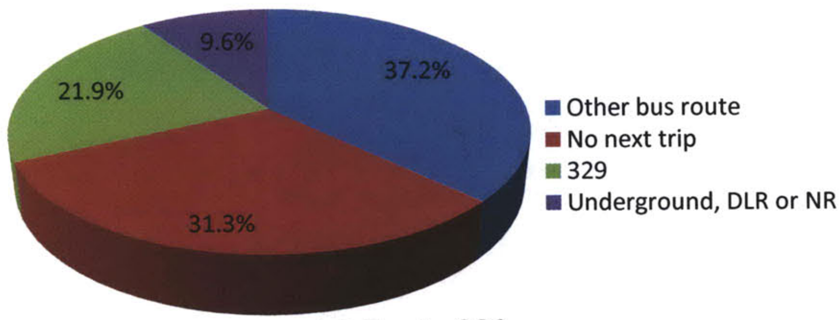
(b) Route 70



(c) Route 185



(d) Route 307



(e) Route 329

Figure 5-17: Subsequent trip travel modes for bus passengers

For example, more than a half of the 24245 Oyster passengers who rode Route 185 on May 12, 2009, took a bus on another route for their subsequent trip while 12.4% reboarded on Route 185, and 6.1% of them took the Underground, DLR or National Rail. The remaining 30.6% of the Oyster passengers did not make any additional trips using their Oyster Smart Cards during the rest of the day (either they only took a single trip that day or were taking their last trip of the day on transit).

A larger percentage of Oyster passengers who ride Route 185 take another bus route (50.8%) for their subsequent trips than those who ride Route W4 (35.4%) and Route 70 (43.3%). On the other hand, around 10% of passengers who ride the W4 or 70 take the Underground, DLR or National Rail for their following trip versus only 6.1% of Oyster passengers who ride Route 185. Route 185 goes through the Central London area, so it has more parallel and intersecting bus routes and thus passengers who ride Route 185 have more access to other bus services. Therefore, it makes sense that more passengers who took Route 185 would take other buses for their subsequent trips than those who took the W4 or 70. Suburban connector Routes 307 and 329 have similar characteristics as Route 70.

The BODS survey actually includes questions asking passengers "How did you get to where you boarded this bus? Another bus, Underground, DLR/Tram, National Rail, walked or other" and "How will you continue your journey after leaving this bus? Another bus, Underground, DLR/Tram, National Rail, walk or other." In the BODS database system, there is an option to show each surveyed passenger's prior and following travel modes, i.e., "Mode to" indicated how did the passenger access the bus and "Mode from" indicated which travel modes the passengers would use when they alighted. Based on the survey results from the BODS database, Tables 5-5 and 5-6 show the travel modes split from BODS survey data and Oyster records, for the subsequent trips of passengers with observed trips on Routes 185 and 329. Note that the data from the BODS survey actually indicate the travel mode for the second leg of a linked trip while no information for the subsequent trips are provided if the subsequent trip is a non-linked trip. One of the biggest problems shown in the following tables is that about half of the surveyed passengers did not answer the question about which mode they would use for their following trip, rendering the analysis for travel modes split of the immediate following trips incomplete and inaccurate. However, with the archived Oyster data, similar analysis could be conducted more easily with less bias and at lower cost.

Table 5-5: Travel modes of subsequent linked trips for passengers on Route 185

Current trip on Route 185	No. of linked trips from BODS	% of expanded BODS ridership	No. of linked trips from Oyster	% of ETM ridership
Next trip by bus	5008	18.6%	8566	32.4%
Rail/Underground/DLR	1756	6.5%	1016	3.8%
Not answered	11763	43.8%	0	
Total expanded ridership	26869		26459	

It is interesting to note that on Route 185, the number of linked bus-to-Underground trips from BODS survey is larger than that from the Oyster records. It is possible that when people filled these survey cards, they might not be sure which travel modes they would take for the immediately following trip, especially for bus routes like Route 185, which runs through the Central London area and has many connections to other bus routes and the rail or Underground or DLR. However, based on the Oyster records, we can obtain information about these passengers' actual travel modes for the immediate following trips.

Table 5-6: Travel modes of subsequent linked trips for passengers on Route 329

Current trip on Route 329	No. of linked trips from BODS	% of expanded BODS ridership	No. of linked trips from Oyster	% of ETM ridership
Next trip by bus	2788	15.5%	5494	28.7%
Rail/Underground/DLR	1055	5.9%	1442	7.5%
Not answered	9214	51.1%	0	
Total expanded ridership	18031		19163	

Most of the passengers on these five selected routes still took buses for the subsequent trips, while only about 6%-11% of passengers would change to the Underground, DLR or National Rail. It seems likely that passengers who took buses would be mostly likely to take buses again for their subsequent trips. This conclusion coincides with the previous research that most of the interchanges within the TfL network are bus-to-bus interchanges. The travel mode split analysis can also be integrated with fare and interchange policy as they can provide more reliable and comprehensive information to transportation planners in TfL.

5.6 Summary

ADC systems generate huge amounts of data. This chapter examines the benefits and potential applications of using these data: they are capable of revealing various aspects with regard to the variability of travel behaviors; they can provide more reliable and comprehensive travel data on each single day for each individual card. The results, which include multi-day trip attributes, demonstrate the possibility and value of applying the ADCS archived data in transit planning. Also, the application of these data can shed more light on evaluation of travel behavior over time and extends the measurement of mobility and service performance to each weekday and weekend. Therefore, transit planners can get a better understanding of the variability of the transit demand and a much clearer picture of ridership patterns, which will also help the transit operators to provide a more customer-oriented supply.

6 Conclusions and Recommendations

This chapter discusses the key results of the bus passengers' travel behavior analysis using ADCS archived data in Section 6.1. It provides recommendations on how to improve the current ADCS in London to get more accurate origin and destination inferences in Section 6.2. Topics for further research using the ADCS archived data for bus network planning are proposed in Section 6.3.

6.1 Summary of Key Results

This research examined the feasibility of using ADCS archived data to analyze bus passengers' travel behavior using London as an example. The first step in this process is to infer the origins for bus passengers by matching the smart card boarding transaction times with the AVL data. It then implements a trip-chaining methodology to infer each bus passenger's alighting location. The origin and destination inference results were then compared with the BODS survey data to validate the automatic inference results with a large-scale survey. This research also develops different approaches to assess and validate the inference results. Finally, this research demonstrates potential applications using the ADCS archived data for bus network planning, with a focus on daily ridership variations and interchange time analysis, and extends the measurement of travel behavior and service performance to each weekday and weekend.

To meet the research objective of examining the potential use of the ADCS archived data to support public transportation service planning, the automatic OD inference methodology was applied to five bus routes, including their connecting routes over a two-week time period. The inferred OD matrices were used to examine daily ridership variation, load profiles and trip distance distributions across weekdays, and between weekdays and weekends. Results show that with the ADCS archived data, more comprehensive and reliable information can be obtained and leveraged to increase planners' understanding of the travel market. BODS survey data were compared with the smartcard transactions on the travel modes used for subsequent trips of bus passengers on the selected bus routes. Finally, further studies on bus passengers' interchange behaviors analyzed the median interchange times on the studied bus routes. The research leads to the following key findings and contributions:

6.1.1 Feasibility and practicality of applying the OD inference method for London buses

The OD inference methodology has been successfully implemented for five bus routes in London, including one of the busiest routes running through the Central London, and two suburban routes. The successful implementation demonstrates the feasibility and practicality of applying the automatic OD inference methodology for London bus routes. As discussed in Chapter 2, as long as the basic information is recorded in the ADC systems, the automatic OD inference methodology can be applied to infer the origins and destinations for bus passengers traveling on any route within the London network.

6.1.2 Variations of daily ridership and travel pattern

Regarding the BODS survey, no information is available regarding the representativeness of the survey day with regard to other days in the period. Smart card data has been used in this research to address this question using a two-week analysis timeframe. This thesis has demonstrated the ability to measure demand and understand its day-to-day dynamics using Oyster smartcard data. Results show that the demand for public transit does indeed vary on a day-to-day basis. Day of week, season, and factors such as weather and events can have a large impact on travel. There is a low probability that any single day of observation will be fully representative of the average day. Ridership varies daily and indeed the location of peak loads can also vary.

Moreover, other bus travel statistics such as passenger-kilometers, average trip length, etc. can be calculated from the inferred OD matrices for any spatial or temporal range of interest.

6.1.3 Bus passenger interchange behavior

As discussed in Chapter 5, based on the archived ADCS data and the inferred OD matrices, the bus passengers' subsequent travel modes and interchange locations and times can also be inferred. Distinguishing linked and unlinked trips as well as the passengers' complete origins and destinations in the transit network is also feasible. This thesis is the first attempt to apply the inferred OD matrices to study interchange behavior. Though only the median interchange times for some selected bus routes are analyzed in this thesis, it is theoretically possible and straightforward to expand similar analysis to the bus-stop level, and to any route-route or route-stop combinations. Such analysis can provide transportation planners more detailed information to help improve the overall bus network performance.

6.2 Recommendations

6.2.1 Implement the proposed methodology to infer bus passenger OD matrices

This research demonstrates the feasibility and ease of applying the trip-chaining methodology to infer bus passengers' boarding and alighting locations. The inferred OD matrices can provide transit planners with more comprehensive and reliable information for service planning than traditional manual surveys. In terms of reduced cost, larger sample size, larger time span coverage and a more automated system, it will be very helpful if London bus planners can implement this methodology for public transportation network planning.

6.2.2 Improve the temporal precision of recorded Oyster transactions

As mentioned in Chapter 3's discussion of the data quality and characteristics, the Oyster transaction times are truncated to the minute while the iBus AVL data are recorded in seconds, which caused difficulties when the two were matched against each other to infer origins. For the short term, the "closest stop" rule is used to match the Oyster transactions to the iBus timestamps that are closest to the Oyster smartcard transaction times. However, if in the future, the Oyster transaction times are recorded in seconds, then the origin inferences should be more accurate. Since the destination inferences were based on the boarding locations of the subsequent trips, improving the accuracy of the origin inferences should also improve the accuracy of the destination inferences to some extent.

6.2.3 Consistent stop naming between BODS and iBus

Section 4.2 discussed the problem that the number of stops, as well as their names and IDs in the BODS and iBus datasets, did not match perfectly. Usually, this mismatch was due to naming discrepancies, such as having a stop named for a landmark rather than a street intersection. There are some notable exceptions on certain routes. For example, on Route W4, BODS uses dummy stops within the "hail-and-ride" segment to pinpoint each passenger's boarding and/or alighting location, while the iBus dataset does not record any stop activity in this segment. In addition to Route W4, such problems were also found on Routes 70 and 307. In this thesis, when validating the origin inference results with the BODS survey, the geographical locations and landmarks for each bus stop recorded in the iBus system were used and some iBus system recorded stops were combined to match certain BODS stops. However, this is very time-consuming and may produce human errors when combining stops and matching the iBus and BODS stops. If the BODS and iBus systems use the same number of stops and consistent names, then no such manual work will be needed. And the origin and destination inference results could be more readily validated.

6.2.4 More complete and accurate iBus information

During the process of inferring bus passenger boarding and alighting locations, one of the biggest problems is a lack of iBus data for some bus trips. For the five selected bus routes, around 10% of the Oyster transactions on these routes could not have their destinations inferred due to missing information from the iBus system that prevented these trips' subsequent boarding locations from being identified. In addition, the timestamps in the iBus system for some bus trips are not always consistent, making it impossible to infer origins for some Oyster transactions. The main reason for missing or wrong information in the iBus system is bus drivers failing to log in on the iBus system at the start of a trip, resulting in wrong bus trip numbers, which are inconsistent with the other records. Though the performances of iBus systems are likely to improve over time, these errors could be significantly reduced by having the system, or supervisors at the bus control center remind drivers to log in on time or to implement a more automatic trip numbering scheme based on time and location or directional headings.

6.3 Potential Further Research

Chapter 5 has demonstrated some possible applications of the ADCS archived data in transportation planning, but there are many more potential topics that could also be studied. This section proposes some directly related topics for further application research:

1) The AFC system continuously monitors transit system boardings, which enables transit agencies to identify travel patterns by daily, weekly, and seasonal cycles. There is also the potential to use the AFC data to derive operational indicators revealing the level of service provided to users on a specific day. Various performance measures could be developed to help operators monitor their networks in greater detail. Given this greater precision on both supply and demand, more effective service planning will likely result.

2) Further research could include the disaggregate analysis of both supply and demand. On the supply side, there is the potential to "optimize" equipment use by analyzing operational data and passenger-load information. By combining these operational performance data, with better demand side data, using straightforward applications such as those described in Chapter 5, it should eventually be possible to improve our understanding of the behavior of public transport users. The analysis of individual user behavior will provide additional information to transit planners on the habits of users: departure times, preferred origins and destinations, preferred routes, etc. Although at present the ultimate origin/destination is not being coded, by extending the OD inference methodology to the full multi-modal network and using the registered Oyster card address information (that is available for a growing number of users), it should be possible to estimate complete passenger origin-destination movements and from there, estimate the modal preferences and path choice parameters given the current and historical quality of service on each route and at each interchange point in the network. Moreover, by linking system usage to home addresses for a sample of users, access behavior can also be better understood, for instance how individuals change their behavior with weather or with the impact of improved customer information systems.

3) By using cluster analysis, different user patterns can be identified and clustered into different groups. Currently, the automatically collected data do not contain information on travel purpose, but by identifying typical temporal patterns of boardings for smartcards of similar classes, it may be possible to partition card users into commuters, students and possibly seniors who travel less than others.

If we keep track of the smartcard number over time, we can analyze the survival rate of Oyster cards and retention of different ticket types, which would provide longitudinal information about the network use and information to inform the fare plan and revenue analysis.

There are many valuable steps on the path to taking full advantage of the rich data sources from the ADC systems. Agencies will eventually better understand both the multi-dimensional performance of public transit services and their impact on travel behavior. More reliable and comprehensive information enables public transport managers and planners to understand both their systems and customers more thoroughly, which may lead to significant changes in the effectiveness and efficiency of public transit services in the long term.

REFERENCES

Zhao, J., Rahbee, A., Wilson, N. H. M. 2007, 'Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems', *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 376-387.

Barry, J., Newhouser, R., Rahbee, A., and Sayeda, S., 2002, 'Origin and Destination Estimation in New York City Using Automated Fare System Data'. *Transportation Research Record* 1817, pp183-187.

Mayor of London 2009, Transport for London – Factsheet, Available at: http://www.tfl.gov.uk/assets/downloads/corporate/transport-for-london-factsheet-july-2009.pdf.

Bagchi, M. & White, P.R. 2004, 'The Potential of Public Transport Smart Card Data', *Transport Policy*, vol. 12, pp. 464-474.

Cui, A. 2006, *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*, Master's Thesis, Massachusetts Institute of Technology.

Buneman, K. 1984, 'Automatic and Passenger-Based Transit Performance Measures', *Transportation Research Record* 992, 23-28.

Jang, W. 2010, 'Travel Time and Transfer Analysis Using Transit Smart Card Data', *Transportation Research Board 2010 Annual Meeting CD-ROM*, Washington, DC.

Trépanier, M., Tranchant, N. and Chapleau, R. 2007, 'Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System', *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*, vol. 11, no. 1, pp. 1-14.

Barry, J., Freimer,R., Slavin, H. 2008, 'Using Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City', *Transportation Research Board 2008 Annual Meeting CD-ROM*, Washington, DC.

Transport for London, Available at: http://www.tfl.gov.uk/corporate/media/newscentre/15260.aspx

Transport for London, Available at: http://www.tfl.gov.uk/tickets/14869.aspx

Transport for London, Available at: http://www.tfl.gov.uk/corporate/projectsandschemes/2323.aspx

Transport for London 2008, *London Travel Report 2007*. Transport for London, London.

Hofmann, M., & O'Mahony, M. 2005, 'Transfer Journey Identification and Analyses from Electronic Fare Collection Data', *IEEE Intelligent Transportation Systems 2005 Proceedings*, Vienna.

Chu, K. K. A. & Chapleau, R. 2008, 'Enriching Archived Smart Card Transaction Data for Transit Demand Modeling', *Transportation Research Board 2008 Annual Meeting CD-ROM*, Washington, DC.

Seaborn, C. 2008, *Application of Smart Card Fare Payment Data to Bus Network Planning in London, UK*, Master's Thesis, Massachusetts Institute of Technology.

Okamura, T., Zhang, J. & Akimasa, F. 2004, 'Finding Behavioral Rules of Urban Public Transport Passengers by Using Boarding Records of Integrated Stored Fare Card System', *Proceedings of the 10th World Conference on Transportation Research*, Istanbul.

Appendix A: OD Inference Rates

As mentioned in Section 3.3.1, a slight discrepancy exists between the Oyster database and the GIS files for the names of each Underground or rail station. The stop codes from the GIS outputs are actually the station location codes, not the same as the "station keys" in the Oyster database (note that the Oyster transactions record the boarding locations for the Underground or rail trips and these locations are described by "station keys"). Therefore, we need to match the two different station naming systems before using the look-up table to infer destinations. We have an Excel file by which we can find the corresponding "station keys" in the Oyster database for the stop codes from the GIS outputs.

Section 3.3.2 noted that the destination inference method is based on trip-chaining, and thus the boarding locations for subsequent trips are also needed. Therefore, origins are inferred for bus passenger trips on both the five selected bus routes and all their related routes. The origin inference rates for those related routes are summarized in Tables A-1, A-2, A-3, A-4 and A-5 (only BODS survey day results are listed here). The origin inference rates for these routes on other days over the studied two weeks are quite close to the origin inference rates on the BODS survey days, as supported by Table A-4.

Note that not all the Oyster transactions on the related routes are included in the study. The criteria that I set up to pick the related Oyster transactions on these routes are as follows:

- For each day, first record all the Card IDs that have transactions on the selected bus routes.
- Then pick up all the transactions that these Card IDs have made during the day, no matter whether they occur on the other bus routes or in rail or Underground system.

Because our destination inference methodology is based on the subsequent trip's boarding information for Oyster passengers, we need to know all the transaction information for these passengers who have once taken the selected bus route during the day. By doing so, we will get a much smaller yet complete sample of Oyster passenger trips on the selected bus routes, which will help facilitate the automatic origin and destination inference process.

Table A-1: Origin inference rates for routes that are parallel to or intersecting Route W4

Route	Origin Inferred	Total Count	% of Origin Inferred
29	27426	31906	86%
34	16123	21444	75%
41	23910	24848	96%
73	27749	31451	88%
102	20909	21916	95%
121	17964	19399	93%
123	20786	21300	98%
141	26413	28894	91%
144	18554	19025	98%
149	24794	25972	95%
184	10011	11812	85%
192	7058	7766	91%
217	7254	7712	94%
221	18290	19400	94%
232	4910	5313	92%
243	34617	37018	94%
279	27202	29920	91%
318	2837	3275	87%
329	18514	18934	98%
349	18000	19739	91%
444	4067	5433	75%
W3	24402	26267	93%
W4	8212	8585	96%
Grand Total	410319	486511	84%

Table A-2: Origin inference rates for routes that are parallel to or intersecting Route 70

Route	Origin Inferred	Total Count	% of Origin Inferred
10	1119	1294	86%
14	1069	1211	88%
148	1234	1456	85%
18	1888	2062	92%
187	633	669	95%
205	401	440	91%
207	2562	2954	87%
211	763	853	89%

220	2501	2569	97%
23	4152	4602	90%
260	1251	1357	92%
266	2496	2575	97%
27	2026	2508	81%
28	2504	2724	92%
283	1496	1895	79%
295	2934	3235	91%
31	2313	2492	93%
316	1829	1942	94%
328	3565	3900	91%
332	458	490	93%
345	456	500	91%
36	1599	1676	95%
360	322	364	88%
414	1019	1189	86%
427	661	743	89%
430	327	353	93%
436	266	310	86%
440	402	466	86%
452	3401	3851	88%
49	3036	3102	98%
52	5528	6430	86%
6	1583	1705	93%
607	1816	1877	97%
7	5933	6507	91%
70	11381	12074	94%
72	1717	1778	97%
74	1012	1050	96%
9	1246	1450	86%
94	2375	2615	91%
95	448	529	85%
C1	860	903	95%
E3	1643	1765	93%
N10	32	34	94%
N11	28	30	93%
N18	84	84	100%
N207	114	120	95%
N31	19	20	95%
N52	74	90	82%
N7	139	155	90%
N9	52	62	84%

Table A-3: Origin inference rates for routes that are parallel to or intersecting Route 185

Route	Origin Inferred	Total Count	% of Origin Inferred
10	122	138	88%
108	239	257	93%

11	323	323	100%
12	902	1030	87%
122	1437	1587	90%
124	496	771	64%
133	689	699	98%
136	1344	1731	77%
137	125	127	98%
14	130	132	98%
148	787	826	95%
155	438	458	95%
156	238	264	90%
159	505	527	95%
16	294	322	91%
160	261	269	97%
170	149	156	95%
171	2211	2266	97%
172	310	331	93%
176	3004	3020	99%
180	331	332	99%
181	1021	1087	94%
185	22794	24245	94%
19	89	90	99%
196	360	373	96%
197	1245	1263	98%
199	744	798	93%
2	963	992	97%
202	916	997	92%
208	1358	1423	95%
21	705	741	95%
211	348	393	88%
22	71	71	100%
225	238	247	96%
24	268	280	96%
261	203	249	81%
273	93	93	100%
284	658	705	93%
3	426	439	97%
321	545	584	93%
333	366	367	99%
336	102	136	75%

36	2886	2896	99%
360	170	243	70%
363	773	798	97%
37	640	704	91%
38	175	177	99%
380	115	141	81%
414	69	74	93%
415	142	144	98%
42	376	411	91%
436	1792	1861	96%
44	161	162	99%
45	863	897	96%
468	597	726	82%
47	957	1178	81%
484	990	1085	91%
507	77	80	96%
52	463	496	93%
59	363	391	93%
63	526	555	95%
68	879	938	94%
73	170	190	89%
74	115	115	100%
77	376	394	95%
78	228	278	82%
8	204	219	93%
82	186	189	98%
87	539	580	93%
88	406	419	97%
89	226	230	98%
9	87	120	73%
C1	142	159	89%
C10	116	117	99%
N136	73	89	82%
N155	38	42	90%
N159	45	46	98%
N171	44	52	84%
N2	10	15	66%
N21	19	19	100%
N3	24	25	96%
N343	21	21	100%

N47	44	44	100%
N63	27	27	100%
N68	23	26	88%
N87	42	43	97%
N89	13	16	81%
P12	254	257	99%
P13	462	464	99%
P4	740	871	85%
P5	148	179	82%

Table A-4: Origin inference rates for all related routes to Route 307[6]

Route	Average Origin Inferred	Average Total Count	Average % of Origin Inferred	BODS-day Origin Inferred	BODS-day Total Count	BODS-day % of Origin Inferred
34	443	521	85.0%	394	559	70%
107	372	387	96.1%	433	435	99%
121	1945	2080	93.5%	2051	2299	89%
184	678	825	82.2%	871	1136	76%
191	1043	1130	92.3%	1245	1306	95%
192	275	304	90.5%	312	342	91%
217	426	444	95.9%	438	484	90%
231	286	305	93.8%	350	352	99%
234	216	226	95.6%	237	247	96%
263	574	596	96.3%	625	653	96%
279	1525	1648	92.5%	1928	2012	96%
298	123	133	92.5%	167	171	98%
299	142	152	93.4%	189	192	98%
307	8728	9171	95.2%	9456	10057	94%
313	347	417	83.2%	405	469	86%
317	256	299	85.6%	263	318	83%
326	539	630	85.6%	667	732	91%
329	523	553	94.6%	701	713	98%
349	787	829	94.9%	828	907	91%
377	183	204	89.7%	182	206	88%
383	122	127	96.1%	108	114	95%
384	284	338	84.0%	342	386	88%
491	247	274	90.1%	312	321	97%
N279	52	60	86.7%	60	64	94%

[6] Average means the average for 14 days over the studied two weeks. Results show that there are no big differences for the origin inference rates between the average of 14 days and the BODS survey day.

W8	623	666	93.5%	706	759	93%

Table 4-5: Origin inference rates for routes that are parallel to or intersecting Route 329

Route	Origin Inferred	Total Count	% of Origin Inferred
102	1090	1105	98%
121	2913	3095	94%
123	532	564	94%
125	738	742	99%
141	2366	2425	97%
144	1013	1047	97%
184	300	356	84%
191	607	619	98%
192	293	336	87%
217	361	363	99%
221	744	788	94%
230	532	542	98%
231	379	382	99%
232	564	630	89%
243	604	952	63%
29	1033	1178	88%
299	161	165	97%
307	525	578	91%
313	145	193	75%
317	160	173	92%
329	17033	17496	97%
34	1106	1233	89%
341	256	264	97%
377	66	88	75%
41	988	1219	81%
444	150	153	98%
629	89	93	95%
67	586	625	94%
N29	131	186	70%
W3	939	1087	86%
W4	175	621	28%
W6	1029	1088	94%
W8	1594	1651	96%
W9	247	337	73%

Table A-6 shows the destination inference results for the other two bus routes that have not been discussed in Section 3.3.3. For Route 307, the destination inference rate is very similar to the other three routes, but the inference rate for Route 329 is much higher. For most of the selected bus routes, 10% of the destinations cannot be inferred due to the lack of iBus information for the boarding locations of the subsequent trips. But for Route 329, only 6.4% of the destinations cannot be inferred due to this reason, meaning that the iBus information is more complete on all the routes related to Route 329 than on those related to the other four selected routes.

Table A-6: Destination inference results on the BODS survey day

Reason	307	329
Total transactions with destinations inferred	69.3%	78.5%
Lack of iBus information	9.3%	6.4%
Next boardings not in 1km buffer area	5.8%	5.9%
Invalid OD	3.5%	2.9%
Single trip	7.3%	4.1%
Lack of directional information	4.8%	2.3%
Total Oyster transactions	10057	17496

Among the five selected routes, Route 307 has the highest ratio of single trips (7.3% compared to 4-5% on the other routes), which may be due to this route's running mostly in a suburban area.

Appendix B: Validation of OD Inferences

As discussed in Section 4.1.2, the comparison results for Routes W4, 70 and 307 are similar to Route 329 in that the majority of the data are still consistent among the three datasets. The exception occurs on Route 185, with more detailed explanations following Table B-4.

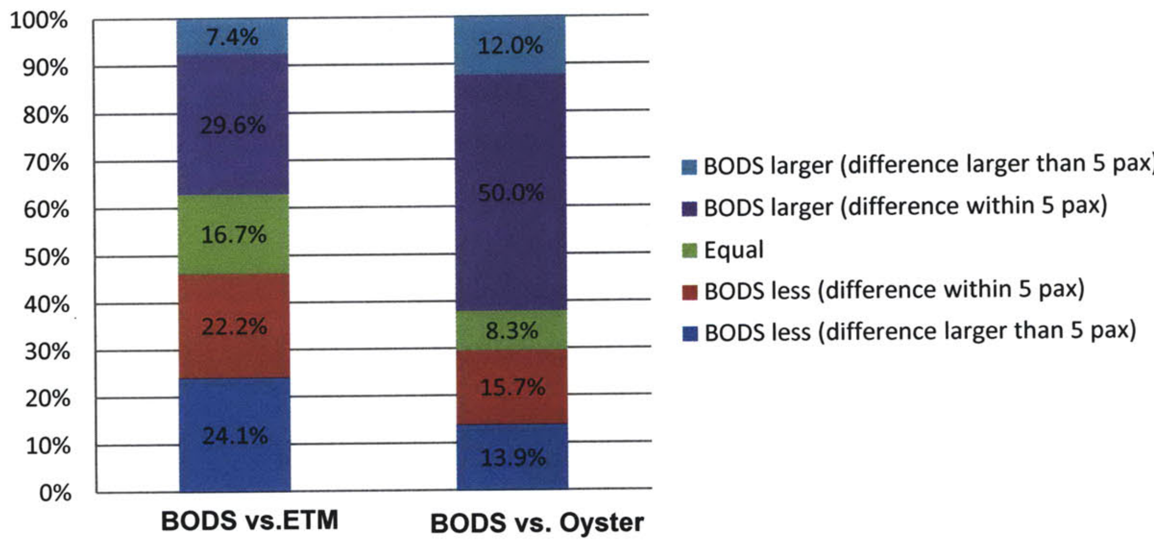


Figure B-1: BODS-ETM-Oyster ridership comparison for Route W4

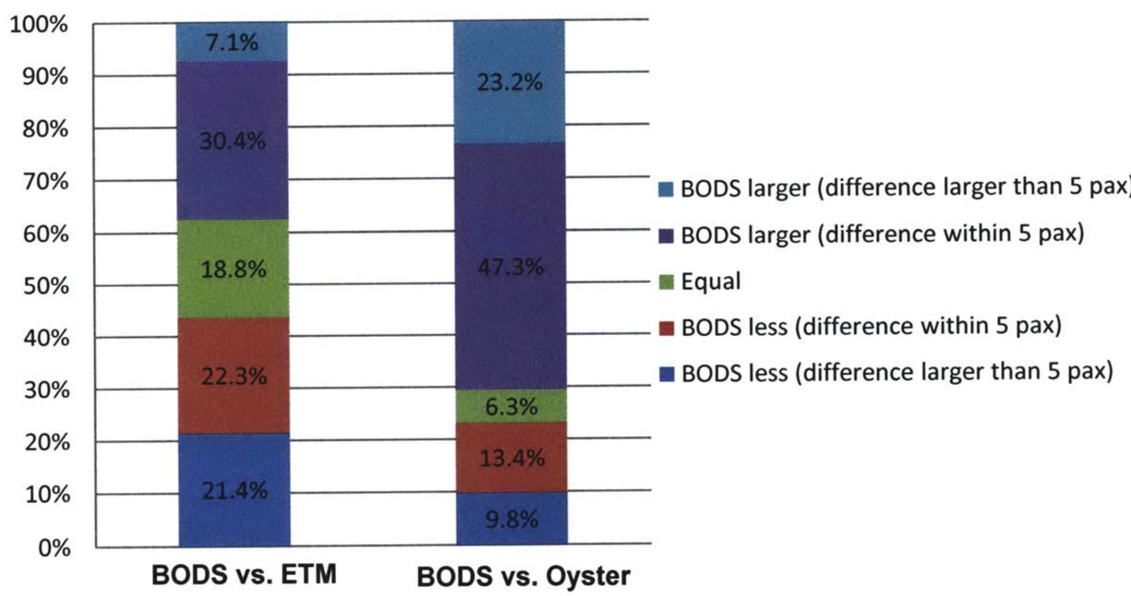


Figure B-2: BODS-ETM-Oyster ridership comparison for Route 70

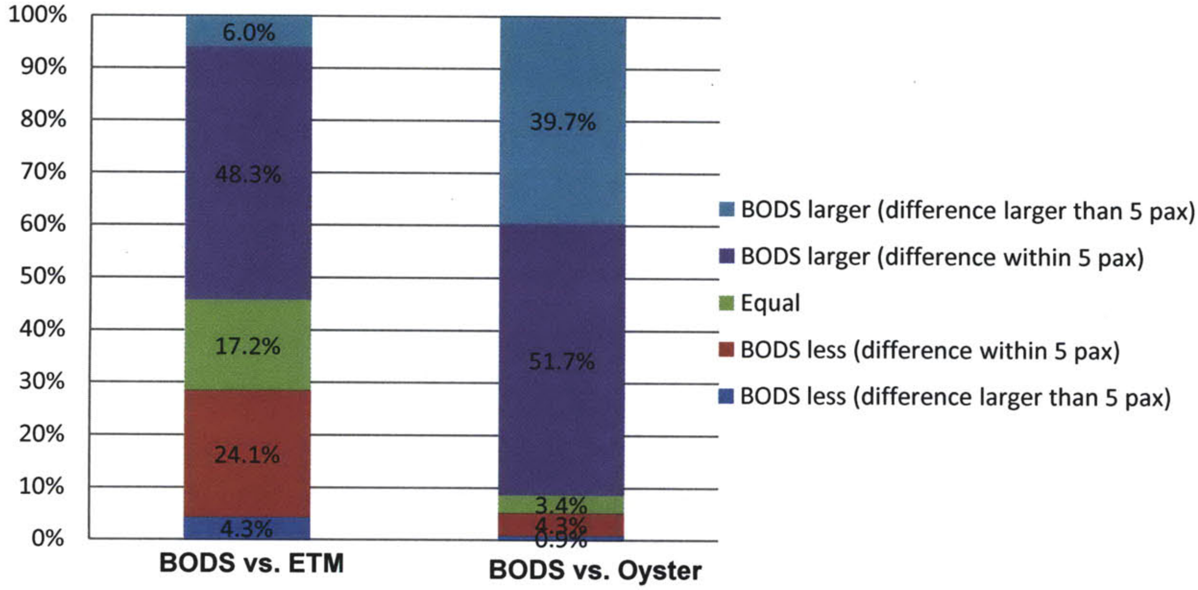


Figure B-3: BODS-ETM-Oyster ridership comparison for Route 307

Route 185 has a special situation, which is partly due to this route's running through the Central London area and being one of the busiest routes in London.

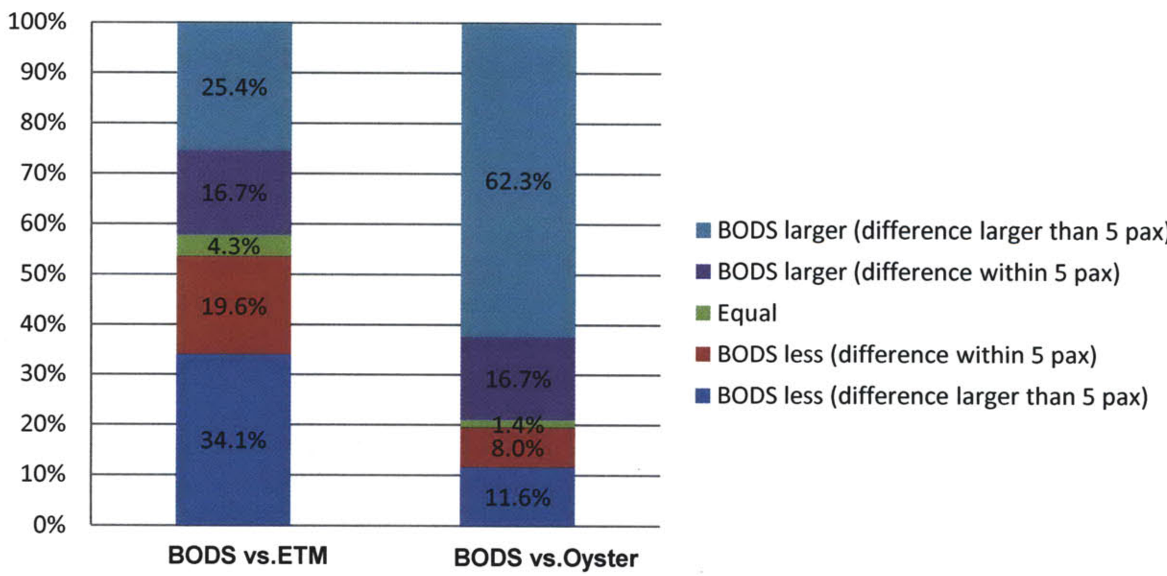


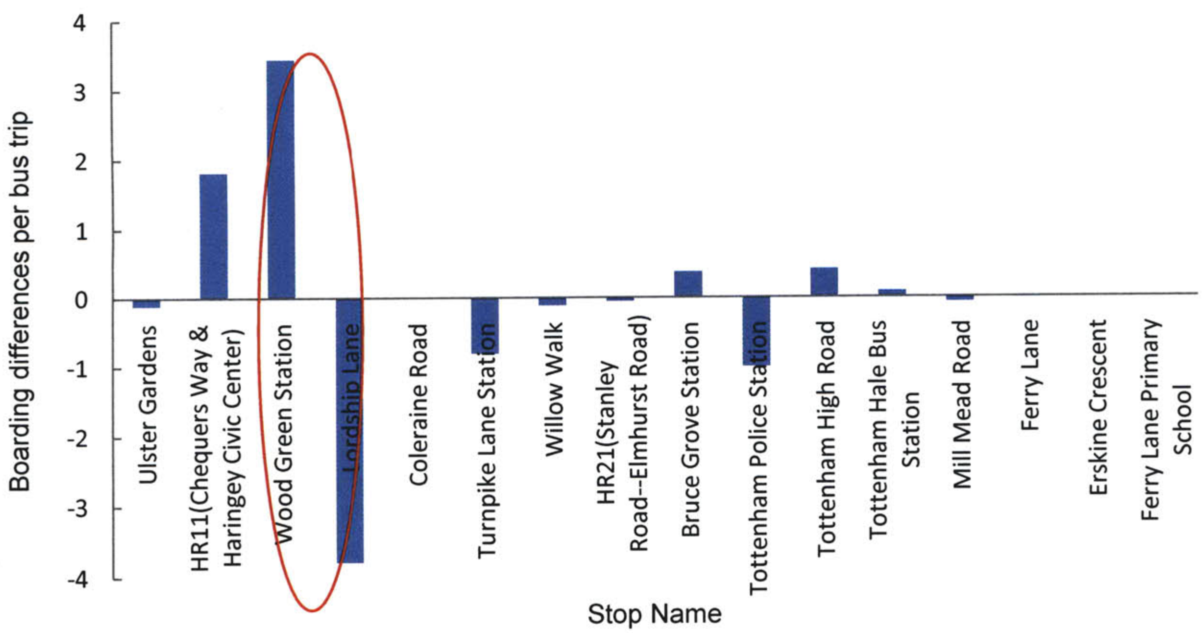
Figure B-4: BODS-ETM-Oyster ridership comparison for Route 185

More surveyed trips on Route 185 have much lower BODS ridership than ETM ridership (34.1%), compared to other routes. This difference may be caused by the BODS survey underestimating the number of boarding passengers for some bus trips during peak hours or some stops in the Central London area. A much larger proportion of the surveyed trips have the Oyster ridership much lower than the BODS ridership (62.3%), meaning there are more non-Oyster passengers for many trips. Since Route 185 has connections to 6 National Rail stations and 4 London Underground stations, it is quite possible that some passengers on Route 185 transferred from these stations do not have Oyster Cards, but just use the paper tickets, especially for some passengers transferring from the National Rail stations, where Oyster Cards were not accepted at all stations in May 2009, when the BODS survey was carried out.

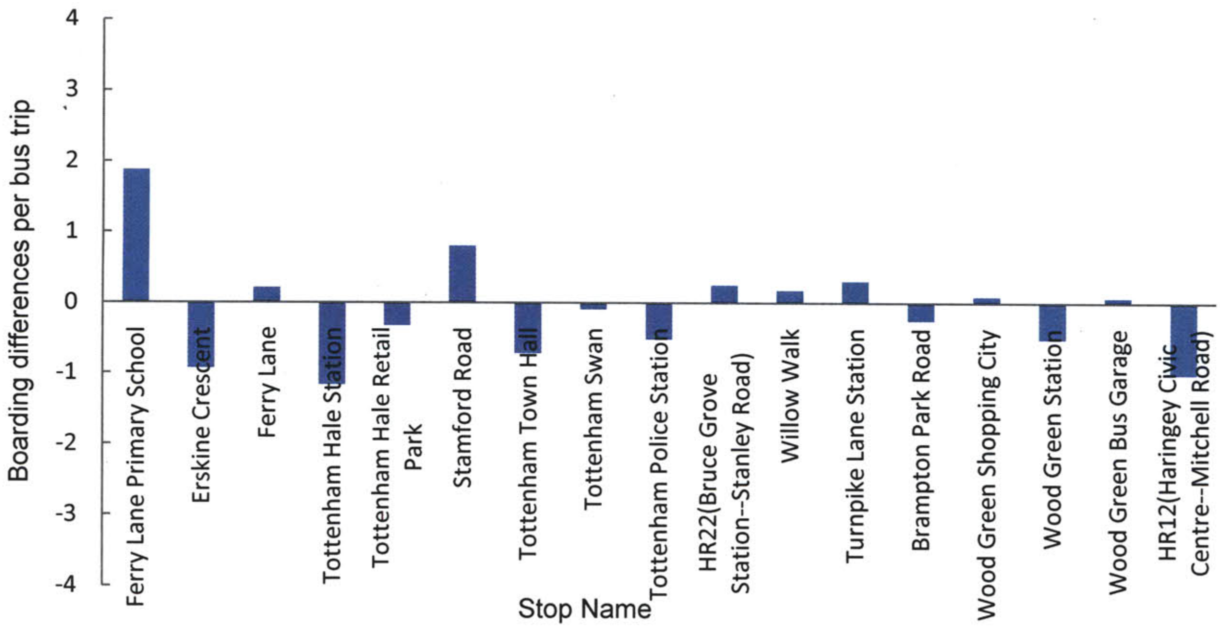
Tables B-1 and B-2, and Figures B-5 and B-6 show the boarding location comparison results for Routes W4 and 70 that have not been covered in Section 4.2. Again, these comparison results support the same conclusion that the number of boardings at stop from the Oyster estimates is close to that from the BODS survey for all the actually surveyed trips.

Table B-1: Summary of boardings from BODS and Oyster (Route W4)

Direction	No. of BODS boardings	No. of Oyster boardings	No. of surveyed bus trips
Eastbound	2374	2368	52
Westbound	2557	2661	56



(a) Southbound

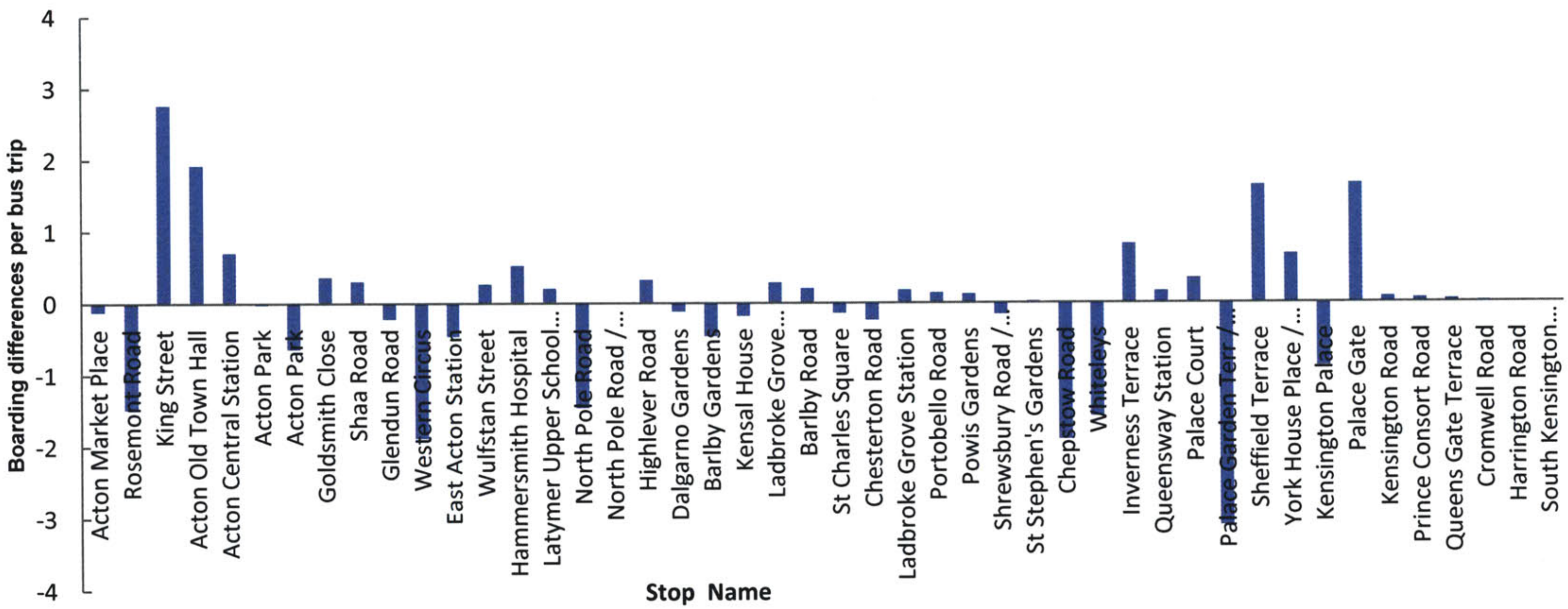


(b) (Northbound)

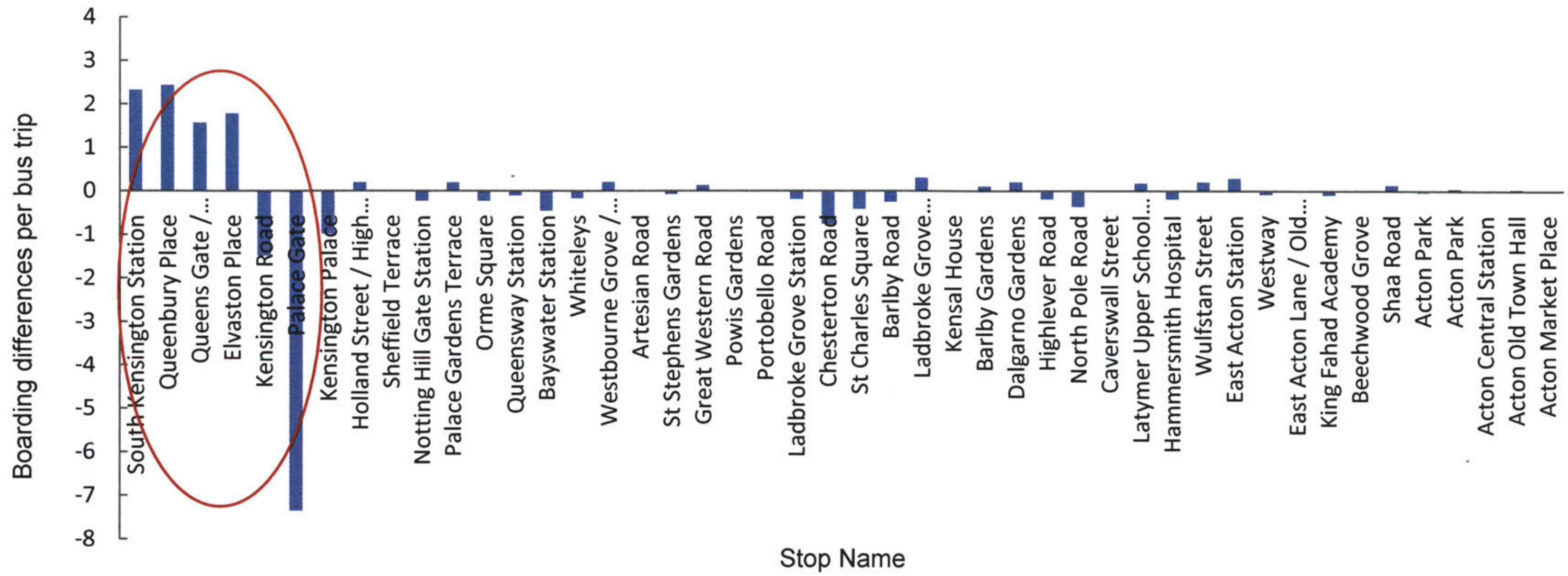
Figure B-5: Boarding location comparison for Route W4

Table B-2: Summary of boardings from BODS and Oyster (Route70)

Direction	No. of BODS boardings	No. of Oyster boardings	No. of surveyed bus trips
Eastbound	3334	3393	50
Westbound	3744	3945	62



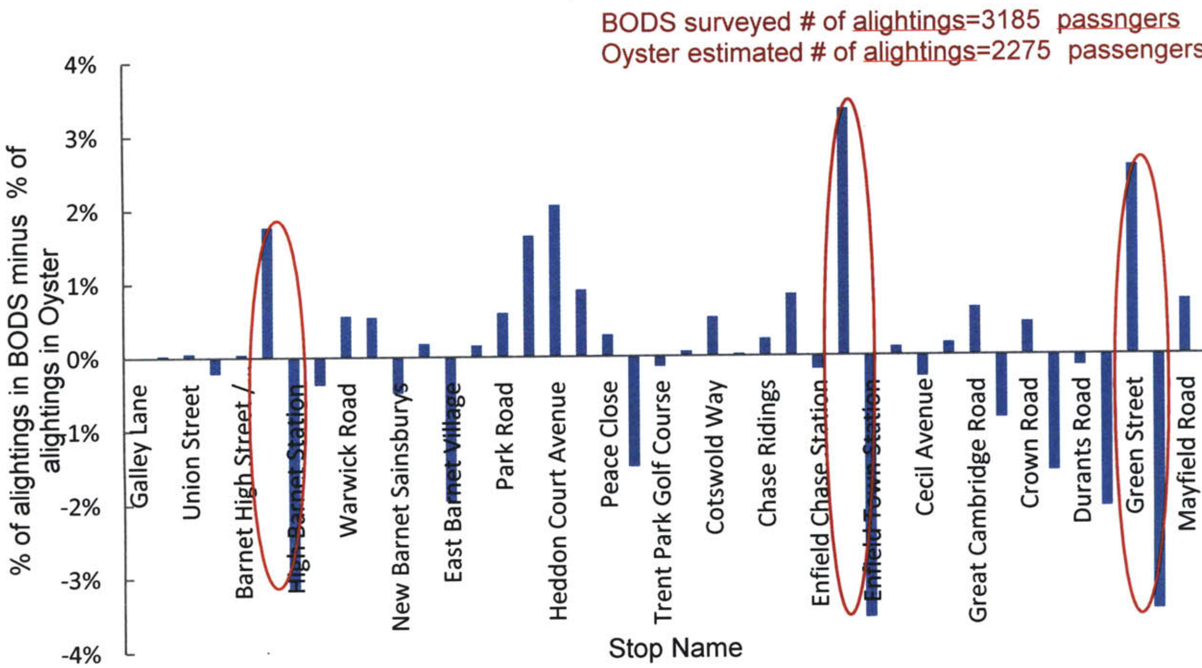
(a) Eastbound



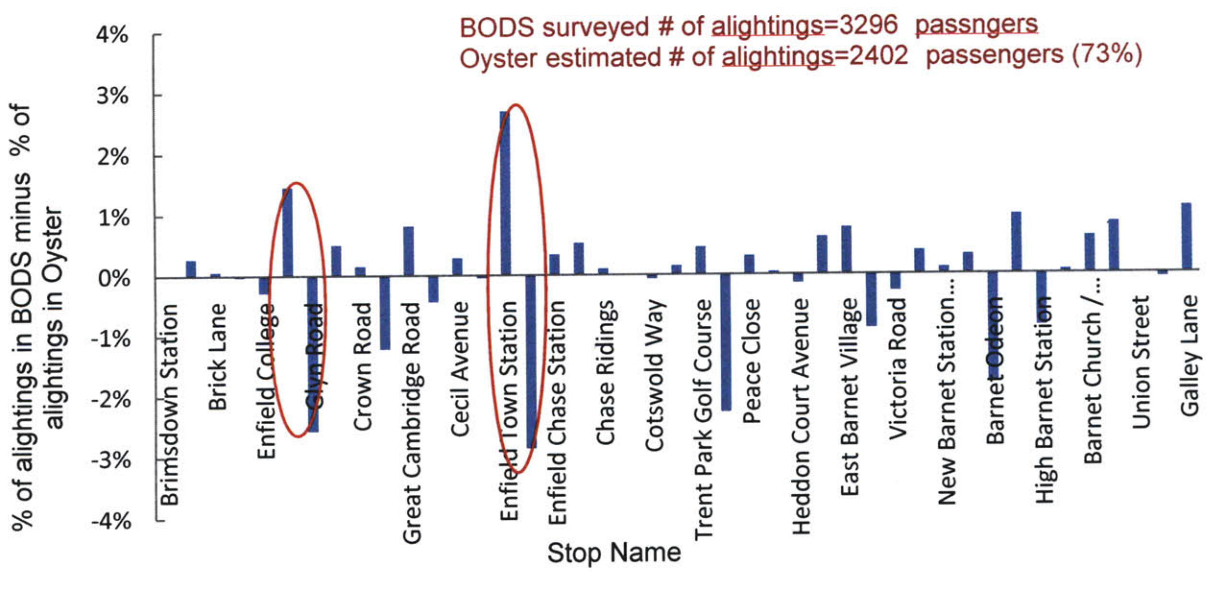
(b) Westbound

Figure B-6: Boarding location comparison for Route 70

Figure B-7 shows the alighting location comparison results as corroborative evidence for Section 4.3.



(a) Eastbound



(b) Westbound

Figure B-7: Alighting location comparison for Route 307

Appendix C: Aggregate Ridership Comparison

Tables C-1, C-2 and C-3 show the ridership comparison results for Routes W4, 70 and 329 that have not been discussed in Section 5.2.3.

Table C-1: Ridership comparison for Route W4

Timeband	Southbound			Northbound		
	BODS	ETM	Oyster/BODS(%)	BODS	ETM	Oyster/BODS(%)
Early AM	93	31	33%	205	82	40%
AM Peak	835	766	92%	854	909	106%
Midday	1977	1920	97%	1819	1793	99%
PM Peak	1054	1235	117%	681	1021	150%
Late	989	986	100%	626	715	114%
All Day	4957	4938	100%	4188	4520	108%

Table C-2: Ridership comparison for Route 70

Timeband	Eastbound			Westbound		
	BODS	ETM	Oyster/BODS(%)	BODS	ETM	Oyster/BODS(%)
Early AM	256	129	50%	97	19	20%
AM Peak	1789	1454	81%	1082	969	90%
Midday	3337	2848	85%	3463	2732	79%
PM Peak	1025	1145	112%	1464	1838	126%
Late	852	690	81%	1173	1219	104%
All Day	7259	6575	91%	7279	5908	81%

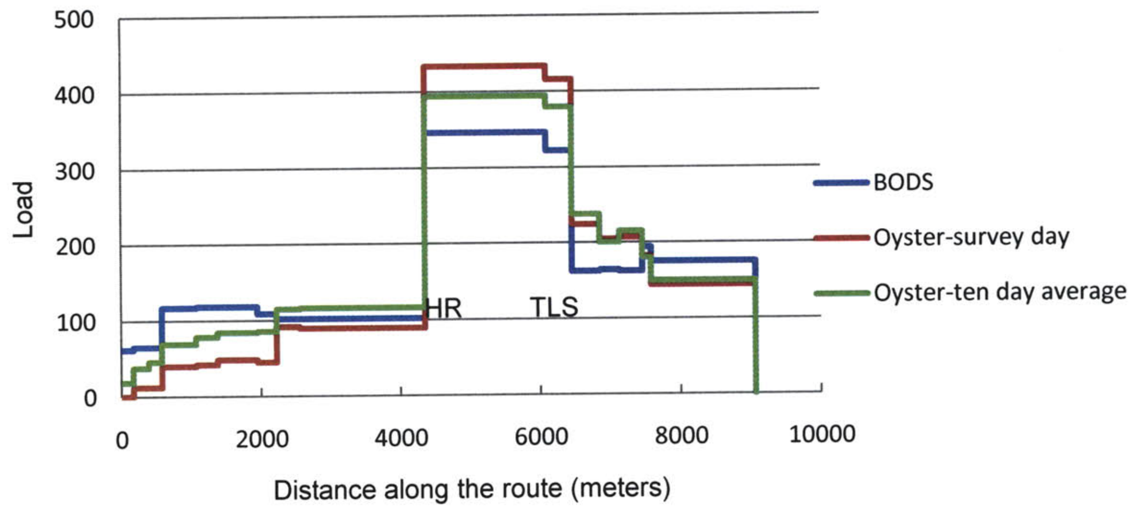
Table C-3: Ridership comparison for Route 329

Timeband	Southbound			Northbound		
	BODS	ETM	Oyster/BODS(%)	BODS	ETM	Oyster/BODS(%)
Early AM	525	407	78%	121	112	93%
AM Peak	1887	1787	95%	1665	1576	95%
Midday	4163	4228	101%	3954	4175	105%
PM Peak	1620	1776	109%	1519	1785	117%
Late	996	1195	120%	1771	2122	120%
All Day	9001	9393	104%	9030	9770	108%

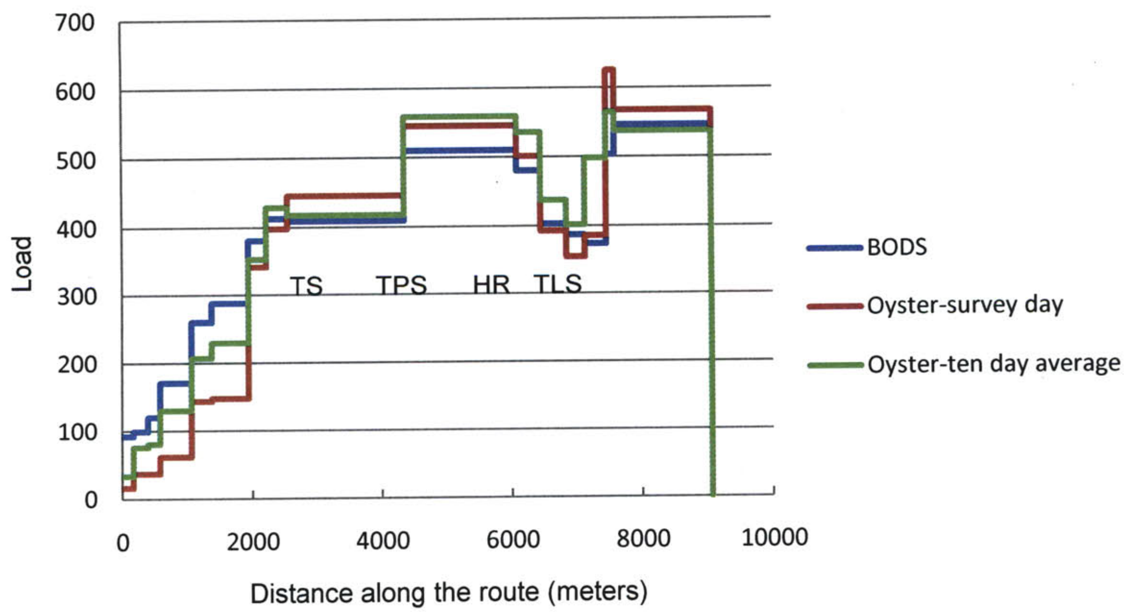
Appendix D: Load Profiles for Peak Directions

The following 15 figures demonstrate the load profiles in the AM Peak, Midday and PM Peak time periods for the five selected bus route in their peak directions.

1. Route W4[7]

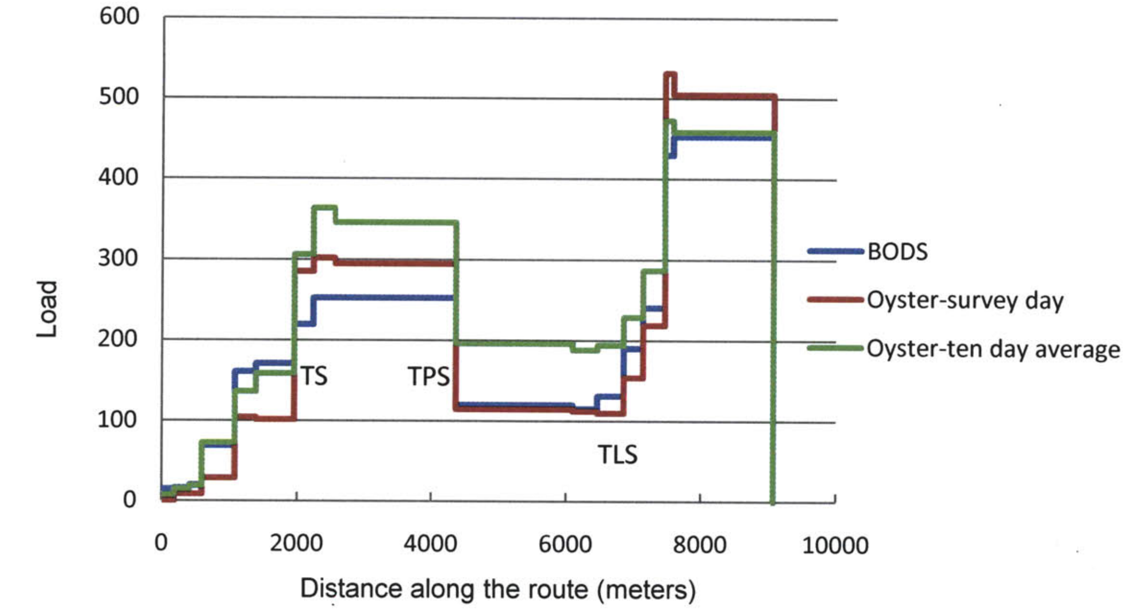


(a) AM Peak (Northbound)



(b) Midday (Northbound)

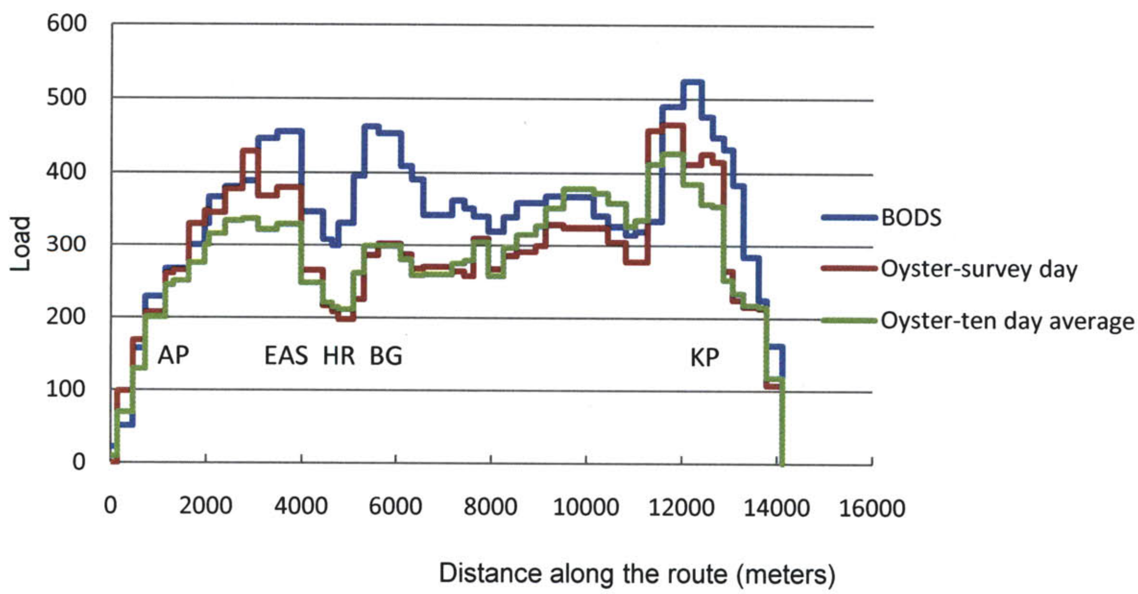
[7] HR=Hail and Ride TLS=Turnpike Lane Station
TS= Tottenham Swan TPS= Tottenham Police Station



(c) PM Peak (Southbound)

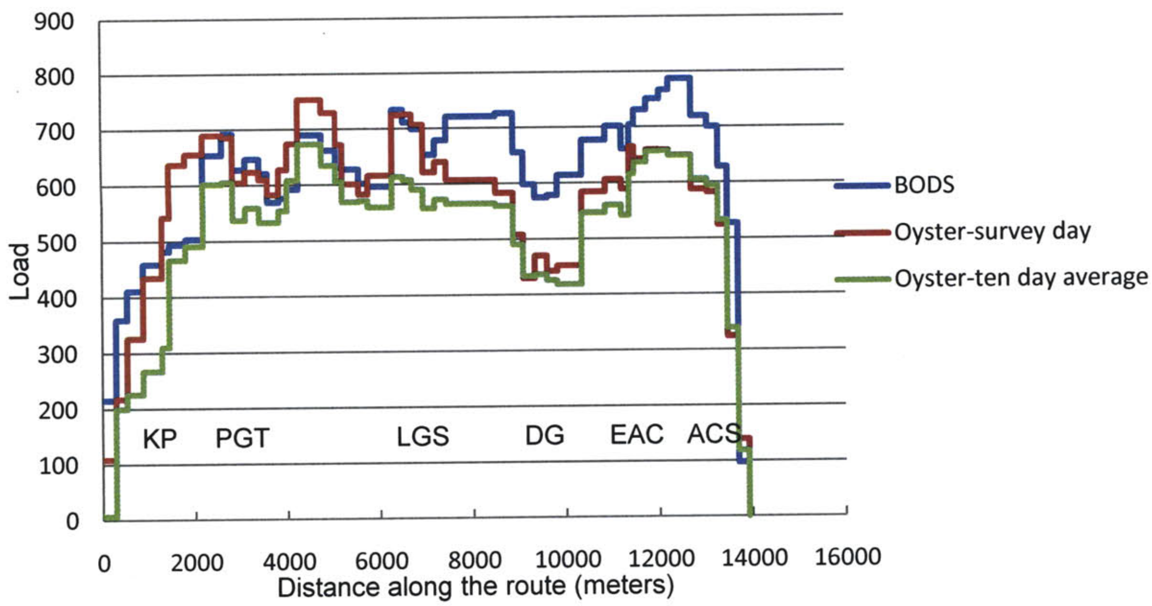
Figure D-1: Load profiles for Route W4

2. Route 70[8]

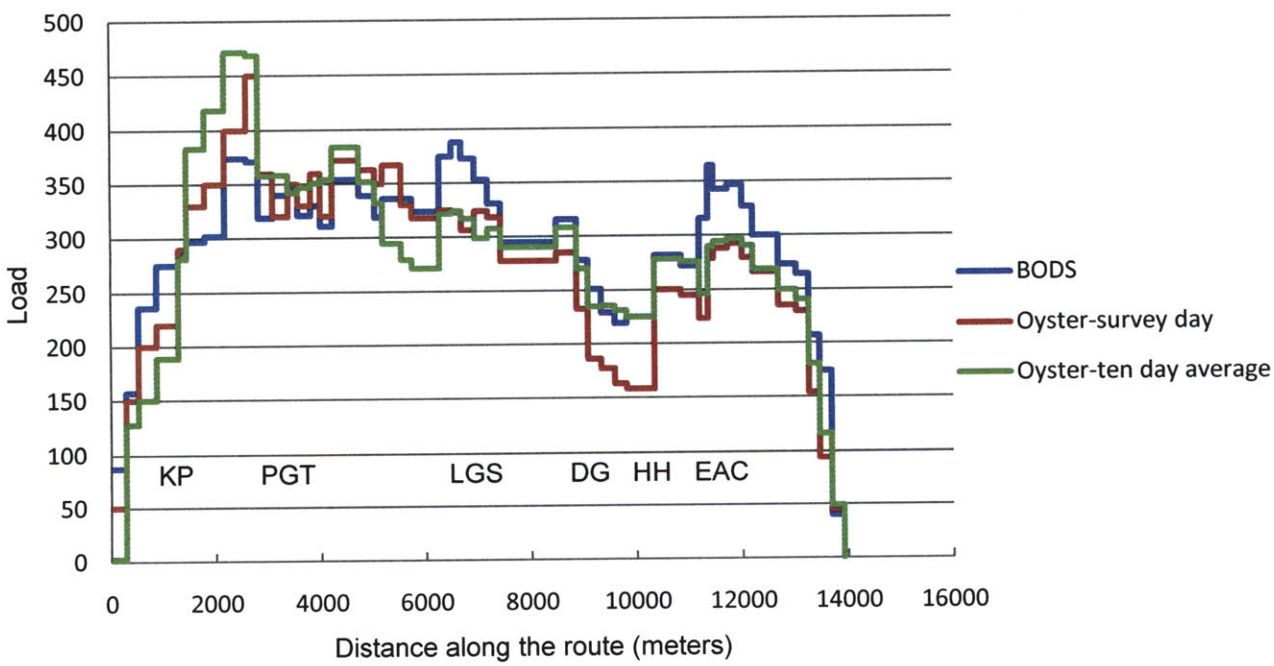


(a) AM Peak (Eastbound)

[8] AP= Acton Park EAS=East Acton Station HR=Highlever Road BG= Barlby Gardens KP= Kensington Palace PGT=Palace Gardens Terrace LGS=Ladbroke Grove Station ACS=Acton Central Station DG= Dalgarno Gardens HH=Hammersmith Hospital EAC=East Acton Savoy Circus



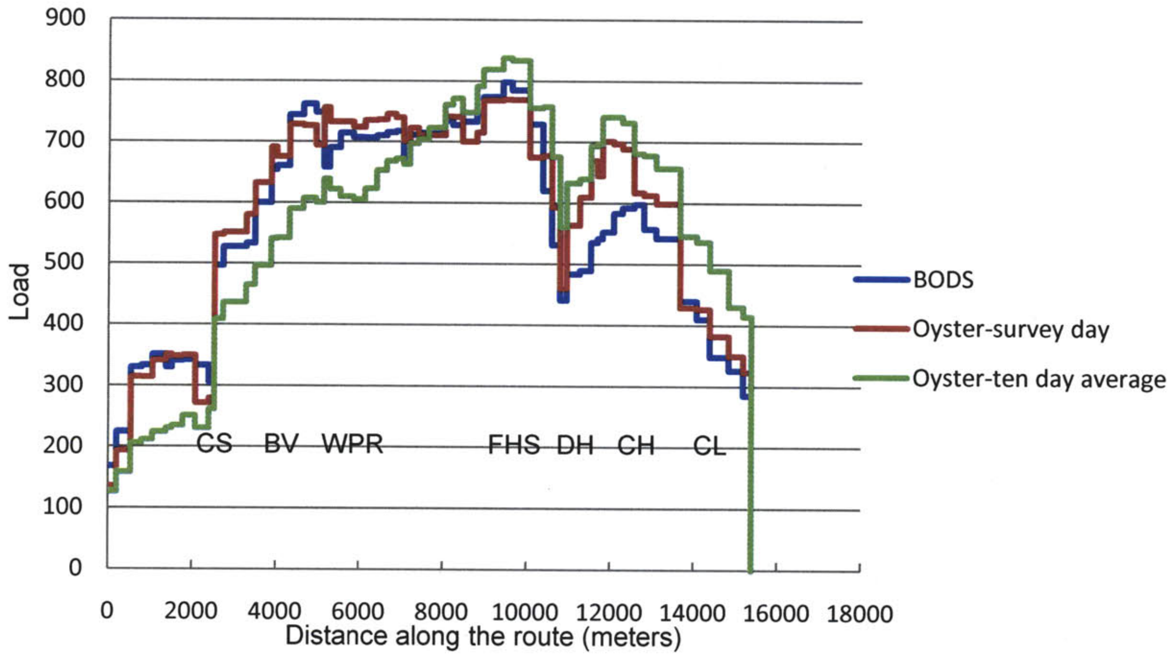
(b) Midday (Westbound)



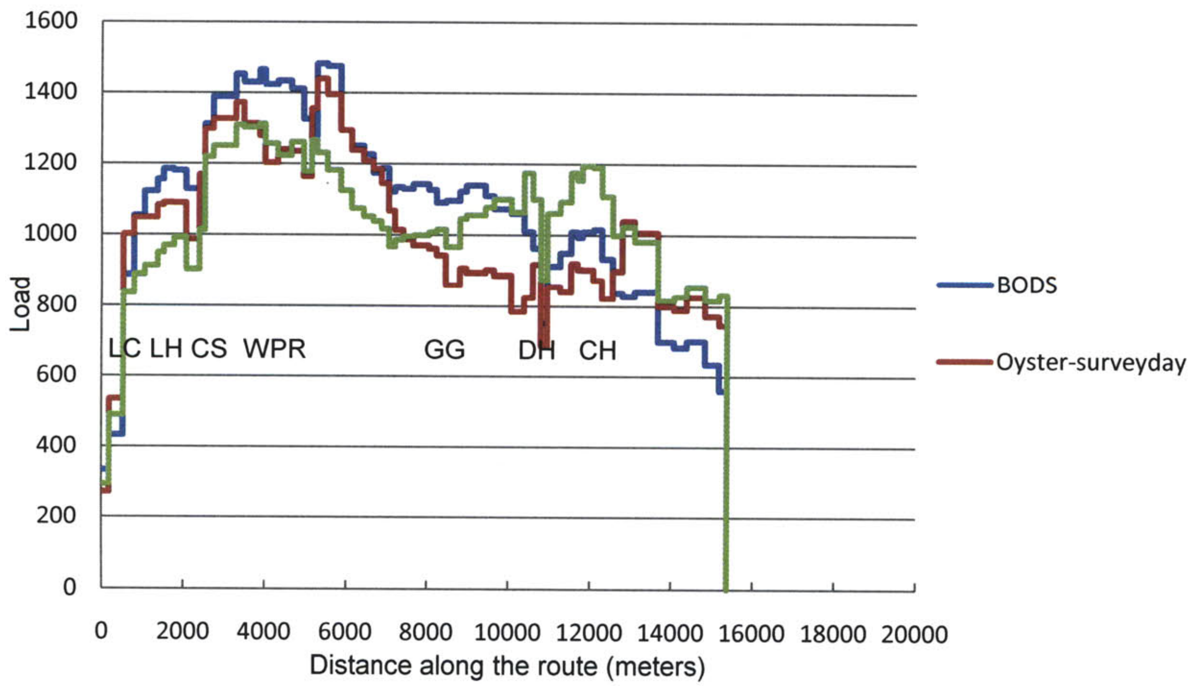
(c) PM Peak (Westbound)

Figure D-2: Load profiles for Route 70

3. Route 185[9]

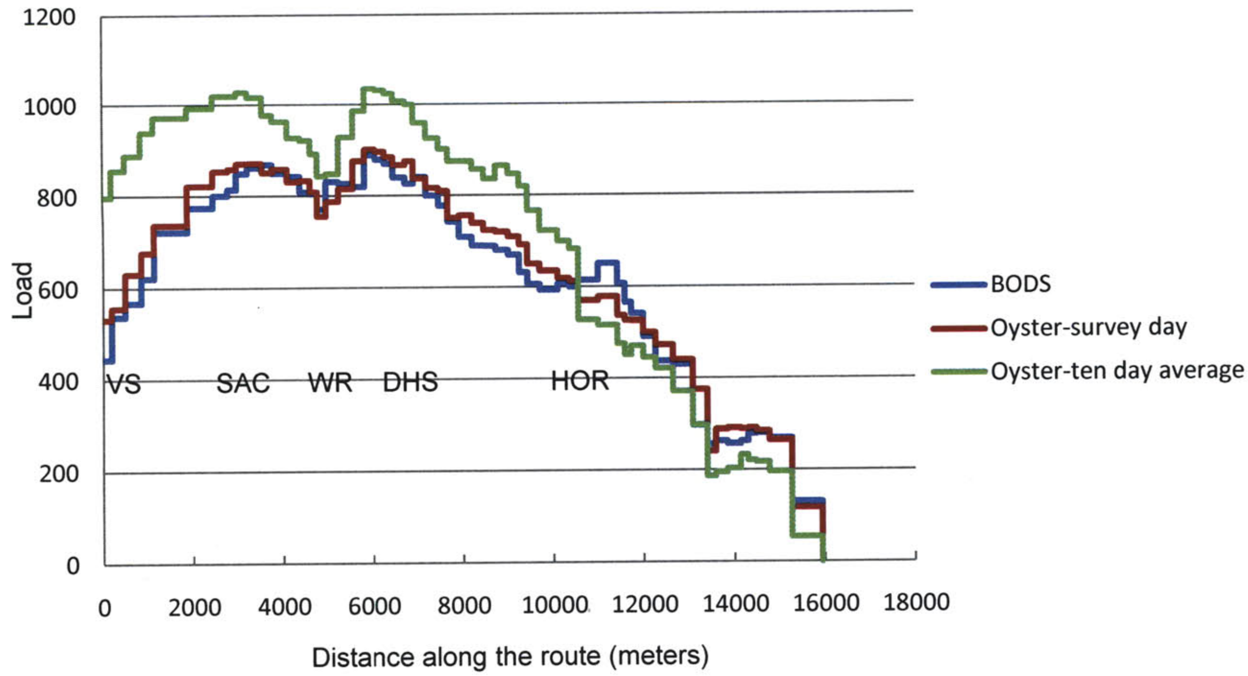


(a) AM Peak (Westbound)



(b) Midday (Westbound)

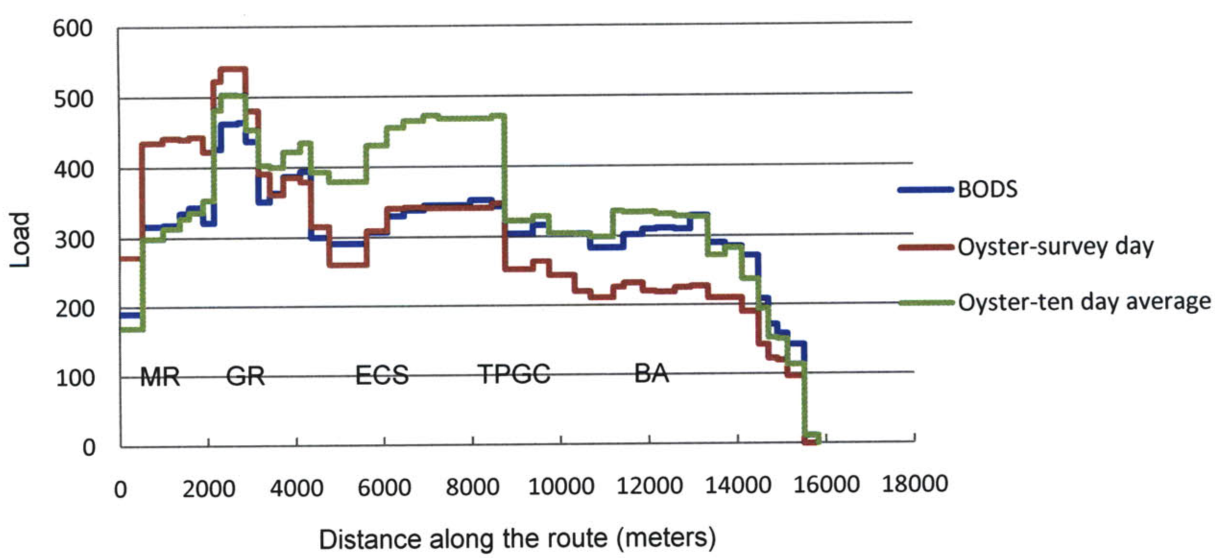
[9] CS=Catford Station BV=Blythe Vale WPR=Waldram Park Road FHS=Forest Hill Station DH= Denmark Hill CH= Champion Hill CL= Coldharbour Lane LC= Lewisham Centre LH= Lewisham Hospital GG= Goose Green VS= Victoria Station SAC= St Anne's Church WR= Warner Road DHS= Denmark Hill Station HOR= Honor Oak Road



(c) PM Peak (Eastbound)

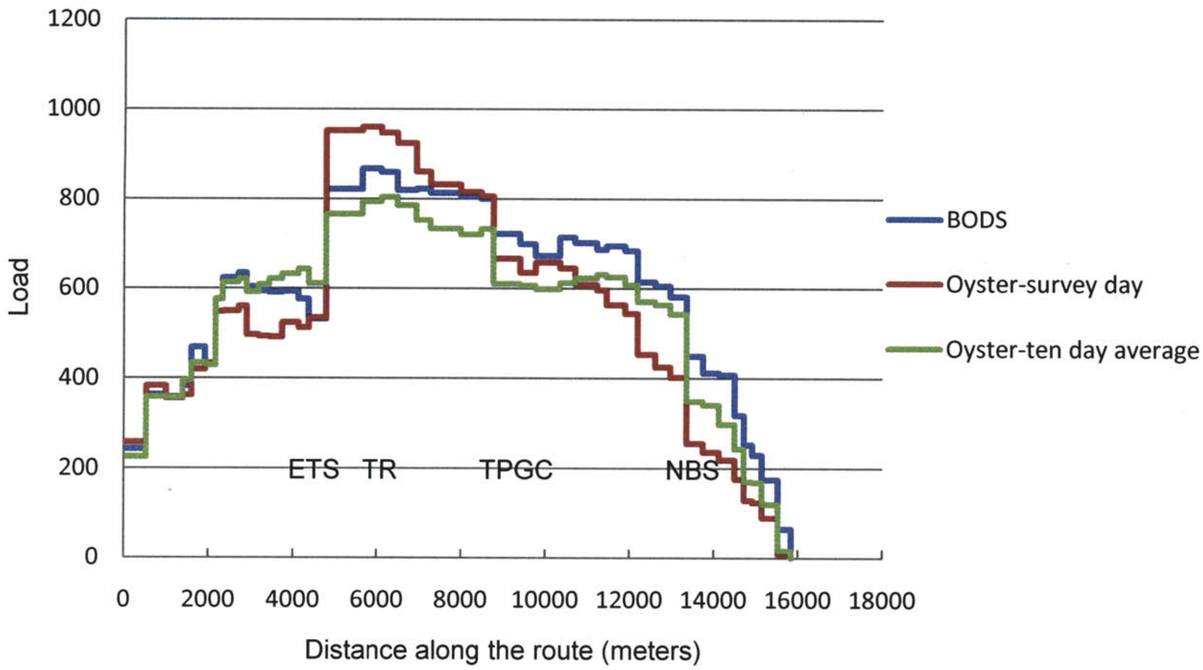
Figure D-3: Load profiles for Route 185

4. Route 307[10]

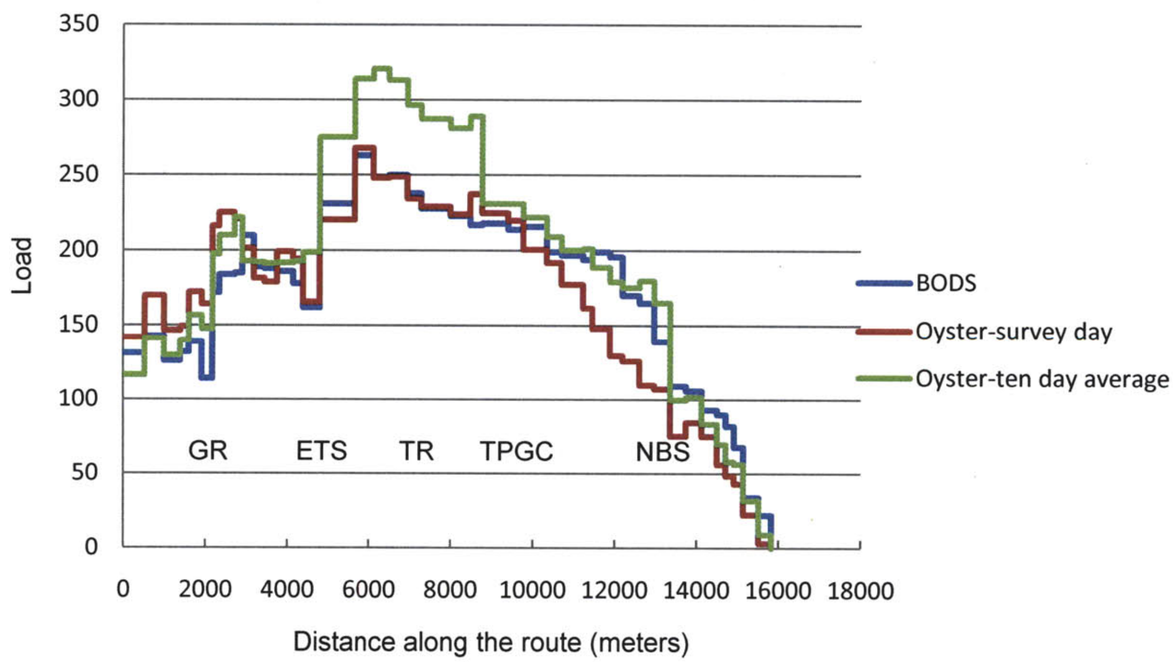


(a) AM Peak (Westbound)

[10] MR= Mayfield Road GR= Glyn Road ECS= Enfield Chase Station TPGC= Trent Park Golf Course BA= Belmont Avenue ETS= Enfield Town Station TR=The Ridgeway NBS=New Barnet Station



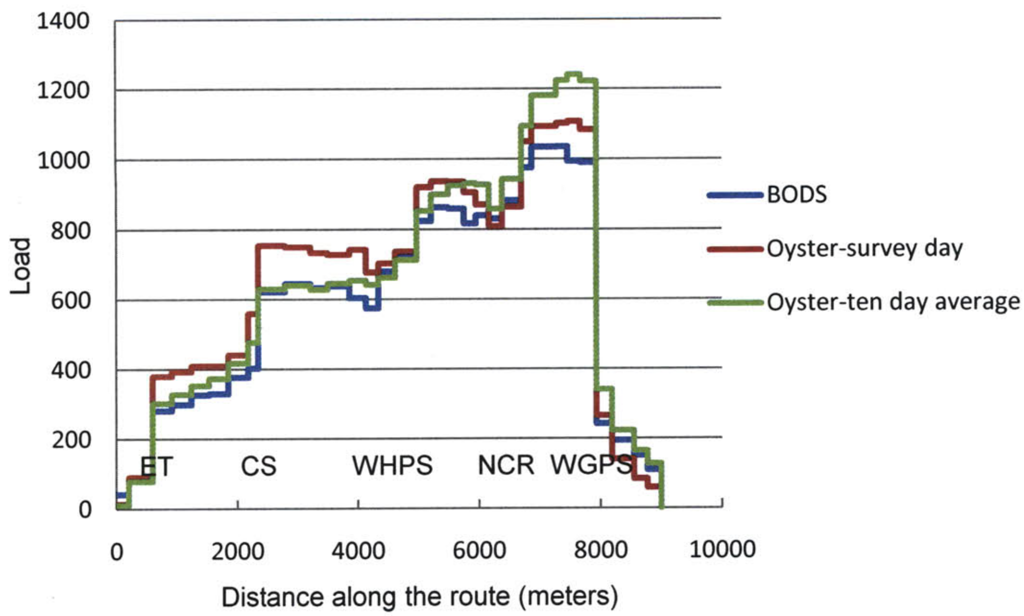
(b) Midday (Eastbound)



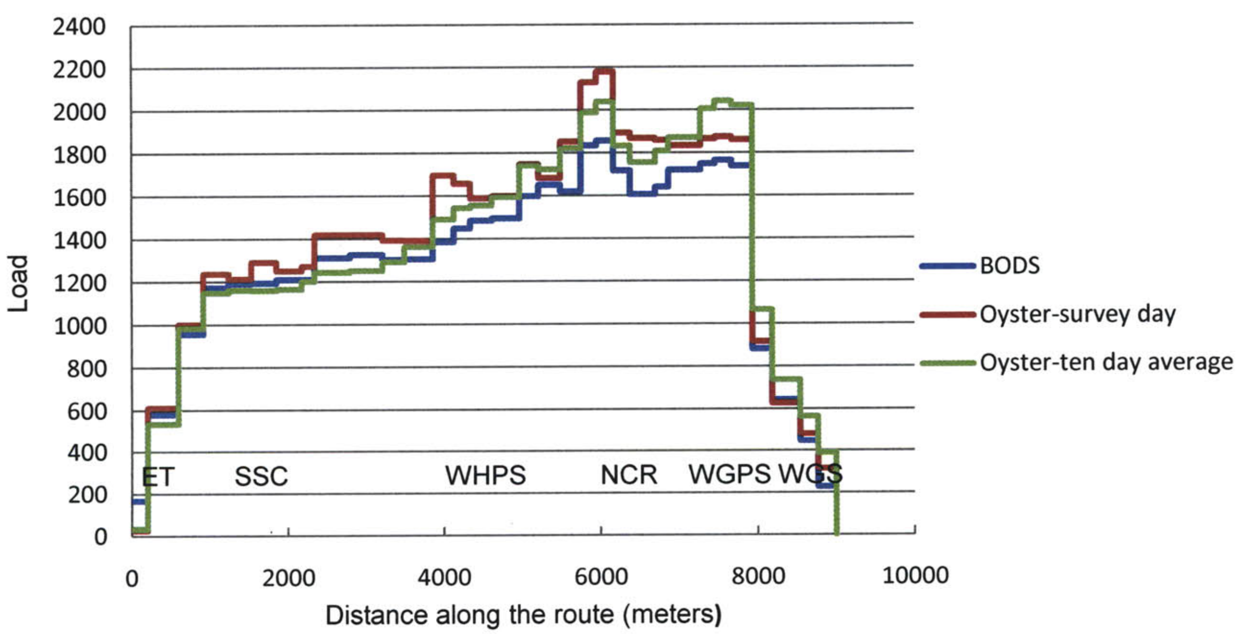
(c) PM Peak (Eastbound)

Figure D-4: Load profiles for Route 307

5. Route 329[11]

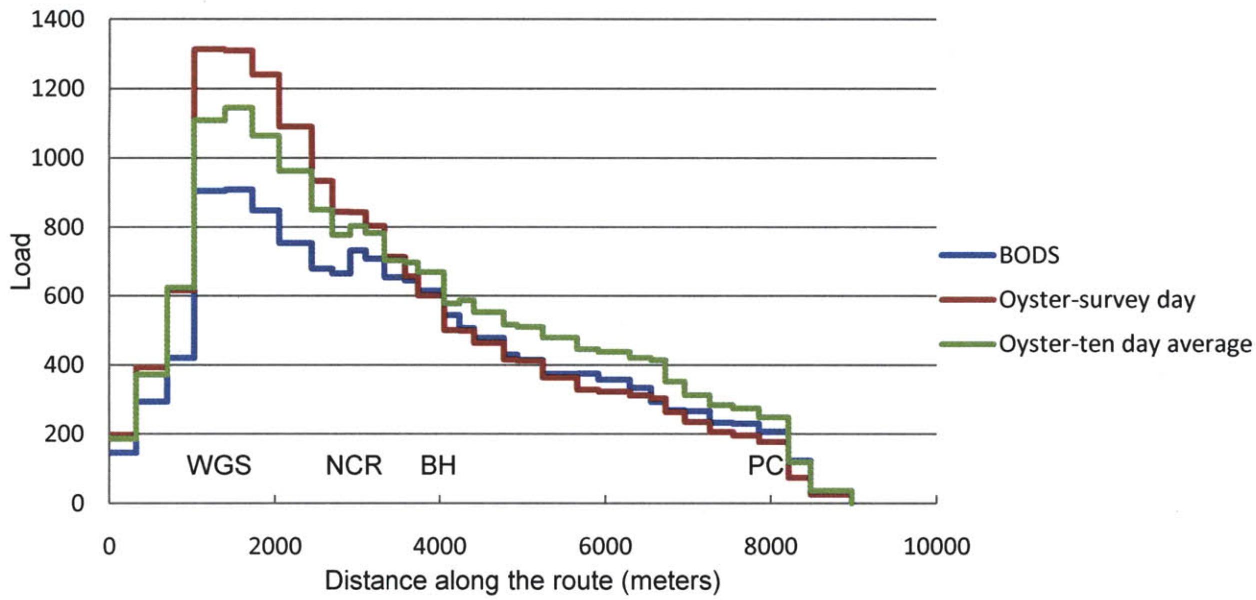


(a) AM Peak (Southbound)



(b) Midday (Southbound)

[11] ET= Enfield Town CS= Church Street WHPS= Winchmore Hill Police Station NCR= North Circular Road WGPS= Wood Green Police Station SSC= Saint Stephens Church WGS=Wood Green Station BH= Bourne Hill PC= Park Crescent



(c) PM Peak (Northbound)

Figure D-5: Load profiles for Route 329