Route-Level Transit Passenger Origin-Destination Trip Estimation from Automatic Passenger Counting Data: A Case Study in Edmonton

by

Cheng Lan

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

Master of Science
in
Transportation Engineering

Department of Civil and Environmental Engineering
University of Alberta

ABSTRACT

Transit passenger origin-destination (OD) trip estimation is very important for transit planning, service management and operation analysis. The traditional method to conduct transit OD trip estimation requires on-board surveys to collect passenger on-off data, which are time-consuming, expensive and usually by-products of other comprehensive censuses which may take place in a very low frequency. The Automatic Data Collection (ADC) systems, including Automatic Vehicle Location (AVL) system, Automatic Passenger Counting (APC) system and Automatic Fare Collection (AFC) system, can collect passenger boarding and alighting counts frequently and have a much larger coverage than on-board surveys. In this thesis, data structure and methods of preprocessing APC data are discussed; route-level transit passenger OD trip estimation methods using APC data are reviewed and applied to the APC data of the Route 1 of the Edmonton Transit System (ETS). The analysis in this thesis shows those methods can produce similar results, but they have strengths and drawbacks. This thesis compares them and makes recommendations for practical applications. Besides, this thesis reviews and implements the stop grouping method to group similar stops along the Route 1 of ETS. The result stop group configuration synthesizes important flow patterns along the Route 1 which is more useful for transit agencies than stop-to-stop OD trip estimations.

ACKNOWLEDGEMENT

This thesis could not have been completed without the help and support of many people, only a few of whom are listed below.

I would like to thank my committee members, Dr. Amy Kim, Dr. Zhi-Jun Qiu and Dr. Mansih Shirgaokar, for their guidance. I am especially grateful to my supervisor, Dr. Qiu, who has always encouraged me to do my best during my two-year program at the University of Alberta. I benefit a lot from his ideas, advices and experiences.

Thanks to my team members in the Centre for Smart Transportation which has a wonderful research and collaboration atmosphere. I would like to thank Dr. Hui Zhang, Dr. Zhen Huang, Gang Liu and Rajib Sikder. Their instruction provides significant support not only on the research ideas, but also on the research techniques. I would also like to thank Dr. Karim EI-Basyouny, Xu Wang, Ying Luo, Xu Han, Lin Shao, Lu Mao, Yahui Ke, Ling Shi, Xiaobin Wang, Qian Fu, Rokib S A and Ran Li, who have always been generous offering their help during my graduate studies.

Also thanks to the Edmonton Transit System who provides APC data of the Route 1 to support this research.

In addition, I wish to express appreciation to my family. Only with their support and understanding can I finally finish my graduation study. Thanks for their loving consideration and great confidence in me through all these years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADC	Automated Data Collection
AFC	Automated Fare Collection
APC	Automated Passenger Counting
APM	Alighting Probability Matrix
AVL	Automatic Vehicle Location
DATS	Disabled Adult Transit Service
EB	eastbound
ETS	Edmonton Transit System
IPF	Iterative Proportional Fitting
IPF-IB	Iterative Proportional Fitting with a Iteratively Improved Base
LRT	Light Rail Train
LRV	Light Rail Vehicle
OD	Origin-Destination
WB	westbound

CHAPTER 1. INTRODUCTION

This chapter presents the background of route-level transit passenger origin-destination (OD) trip estimations from Automatic Passenger Counting (APC) data. It also indicates problems of the traditional on-board survey method, describes the research motivation, research objectives and the structure of this thesis.

1.1 Background

The origin-destination (OD) trip estimation, which is also called "OD matrix", "trip matrix" [1] or "trip table" [2] [3], reflects trips from origins to destinations during a time period. Origins and destinations are geographical areas defined by factors like transportation network details, land use, population and potential growth in future. The OD trip estimation reveals travel patterns within study areas. Therefore it is required in transportation planning and management.

The OD trip estimation used to figure out passenger flows between bus stop pairs or stop group pairs is referred to as transit passenger OD trip estimation. It can be adopted to analyze a transit network or a single route. It helps to conduct transit ridership estimation, headways setting, before-and-after analysis for operation alternatives and so forth. Therefore, the transit passenger OD trip estimation is a critical tool for transit planning, service management and operation analysis [4] [5] [6]. This thesis focuses on the route-level transit passenger OD trip estimation which is used to analyze a single route.

The traditional method to conduct route-level transit passenger OD trip estimations is the on-board survey [3]. Passengers on studied routes may be asked to fill their origin and destination information in questionnaires [4] or they are handed a card with a stop label when boarding and return to researchers when alighting from the bus [7].

With the rapid development of the data acquisition technology, on-vehicle Automated Data Collection (ADC) systems, including Automatic Vehicle Location (AVL), Automatic Passenger Counting (APC) and Automatic Fare Collection (AFC), are widely used in transit buses around the world. AVL records vehicles' location information so that transit agencies can track their positions. APC records the number of boarding and alighting passengers at stops. This technique can work with AVL data together to identify passenger boarding and alighting counts per stop. AFC systems, which are used more and more all around the world nowadays [8] [9], are employed to automate the ticketing system of the public transportation. Data obtained by AFC systems is different from APC systems. It does not record passenger on-off counts directly. But valuable information, including OD trip estimations, is still possible to be derived after processing. Cui [10] summarized and discussed the basic idea about how to estimate transit OD trips from AFC data. Munizaga and Palma developed this idea further and applied it to the multimodal public transportation OD trip estimation [9].

1.2 Problem Statement

The traditional method to produce transit passenger OD trip estimations relies on the on-board survey. It is time-consuming and manpower-intensive. Personnel have to distribute the questionnaires to passengers onboard and spend long time to process them. Hence this kind of surveys is usually very expensive and difficult to implement, especially for large-scale sampling [1] [11]. Furthermore, collecting passenger on-off counts usually is a by-product of occasional system-wide on-board passenger surveys in fact [10]. It is rarely to conduct on-board surveys to collect passenger on-off data only for the purpose of transit passenger OD trip estimations.

Besides, on-board surveys may result in small sampling size. Because of those economic constraints, only several bus trips or limited time periods may be observed and small portion of passengers may be surveyed. Moreover, not all passengers may response the on-board survey. Researchers noticed passengers making short trips tend not to respond to the questionnaires [12]. The OD trip estimation thus derived from the on-board survey is tend to be underspecified and cause firm biases for transit agencies to make various decisions [10].

1.3 Research Motivation

The Edmonton Transit System (ETS) does not conduct transit passenger OD trip estimations from data collected by either traditional on-board surveys or ADC for some reasons. However, they have a large amount of boarding and

alighting counts collected by the APC system. These counts reflect passenger volumes at different bus stops and can be used to estimate transit OD trips.

This thesis is going to find a practical way to conduct route-level transit passenger OD trip estimations from the APC data provided by ETS and figure out route-level transit trip patterns. Researches about AFC data are promising and are very good complements for transit OD trip estimations from APC data [9] [10], but they are not included in this thesis since the Edmonton Smart Fare System project is ongoing and has not finished yet at this point [13].

1.4 Research Objectives

This research proposes a practical procedure to conduct the route-level transit passenger OD trip estimation between stop groups from the APC data provided by ETS. There are three specific goals of this thesis:

1) Develop APC data preprocessing methods to process large amount of stop-based APC data provided by ETS.

2) Review different route-level transit passenger OD trip estimation methods, apply them to ETS's APC data, analyze results and identify the most suitable one in the case study; and

3) Review and implement a stop grouping method to group similar stops to synthesize important flow patterns along the route in the case study.

1.5 Structure of Thesis

This thesis includes 5 chapters:

Chapter 1 introduces the background of transit OD trip estimations from APC data as well as the problem statement and research objectives.

Chapter 2 is the literature review chapter, which reviews APC system, APC data preprocessing methods, methods to produce route-level transit OD trip estimations and the method to find out stop group configurations.

Chapter 3 describes method implementations in this case study on the Route 1 of ETS.

Chapter 4 analyzes and discusses the results of data preprocessing for APC data of the Route 1, four different estimation methods and the stop grouping method.

Chapter 5 presents the conclusion and provides suggestions for the future works of transit passenger OD trip estimations from APC data.

CHAPTER 2. LITERATURE REVIEW

This chapter summarizes the existing researches about data preprocessing of APC Data for transit OD trip estimations and techniques used to produce transit passenger OD trip estimations for a single route from APC data. The idea about grouping stops is also reviewed in this chapter.

2.1 APC Data Preprocessing

2.1.1 APC System

APC Systems are mainly utilized to collect the number of boarding and alighting passengers for a bus at every stop. Besides the number of how many passengers get on and off at each stop, time stamp and stop location or stop index are also recorded by APC systems.

APC systems can be implemented with different technologies. Four popular technologies [10] are listed as below:

- Infrared Light Beams

A group of two infrared sensors are installed at the same height level, for example, above the doorways to a vehicle. These two sensors emit two spaced infrared light beams. So the order in which a passenger breaks these two beams determines him or her boarding or alighting. It is the most widely used APC technology, and its accuracy is quite high if the passenger volume is not extremely high.

- Pressure Sensitive Mat

The pressure sensitive mat replaces common treads on the bus step wells and detects passengers when they cross bus steps according to the pressure of their feet. However, this technology is easy to be damaged because of foot traffic, water and exposure to the elements. Moreover, this technology is not reliable in buses whose floor is level.

- Passive Infrared Sensors

The passive infrared sensors count passengers by detecting the difference of passenger body temperature and the environment temperature. This technology is sensitive to sudden temperature and light changes and may be under counting due to immobile passengers.

- Cameras

Counting passengers by using cameras equipped inside buses is indeed a process of image recognition of passengers. The accuracy of this technology highly depends on ambient light, height and quality of images taken by cameras.

2.1.2 Potential Problems in APC Data

A trip, in which a bus runs from one terminal to another, should have the same total boarding and alighting counts. However, data collected by APC systems in practice may have different sums of on and off counts.

First of all, the imbalance situation may be caused by devices faults, high passenger volume which is beyond the capability of APC systems, or blind spots

where passengers are not detected when passing. It is indicated in Cui's observation that APC systems are easy to undercount passengers [10] .

Secondly, the imbalance situation may be caused by a situation called "passenger carry-over" [4] [13]. Passengers who take buses at stops in the opposite direction which are close to the terminals are most likely to save some walking or secure a seat. The APC system records their boarding but doesn't record their alighting at the last stop.

Besides the imbalance of boarding and alighting counts, negative load errors are also observed in APC data processing [4] [14]. Even the sums of boarding and alighting counts equal to each other, there are some errors in APC data which lead to a situation that the number of alighting passengers at a certain stop is greater than the number of passengers onboard. This problem may be because some passengers or operators get on and then get off immediately. Lu described this problem in a numeric way and Table 1 illustrates this problem [4]:

Table 1 Negative Load Example [4]

Stop Sequence	Observed boarding	Observed alighting	Calculated arrival load	Calculated through load
1	1	0	0	0
2	0	0	1	1
…	…	…	…	…
18	0	0	1	1
19	*2*	*2*	*1*	*-1*
20	1	0	1	1
21	0	0	2	2
…	…	…	…	…
Total	**28**	**30**		

Calculated arrival load in the table above means how many passengers onboard enter a stop, and calculated through load means how many passengers onboard don't alight at that stop. Those who get on at that stop are not counted in the calculated through load. The calculated through load can be calculated from the following equation:

$$L_{through} = L_{arrival} - Q_{alighting} \quad (2\text{-}1)$$

,where

$L_{through}$ = the calculated through load,

$L_{arrival}$ = the calculated arrival load,

$Q_{alighting}$ = the observed alighting counts.

If the calculated through load at a stop is negative, the negative load problem arises. This APC error could break down OD estimation methods which are reviewed in the section 2.2 [4].

At the stop 19 in the table above, there is one passenger enters the stop, but two passengers are observed to alight. It should be impossible. One reasonable explanation might be one passenger gets on and then gets off immediately if there is no device fault.

2.1.3 Data Preprocessing

In order to deal with the imbalance errors in APC, Cui [10] proposed a method which matches lower counts to the higher counts because he argued that APC systems tend to undercount passengers. In this method, the lower count, no

matter it is the total boarding count or the total alighting count, is adjusted to the higher one. The ratio of the observed total count to the adjusted count is calculated as a factor. Then the observed counts of all stops are increased proportionally according to that factor.

Lu mentioned a similar method to deal with the imbalance issue [4]. Instead of matching to the higher count, this author considers to match the average of the sum of boarding counts and the sum of alighting counts. This method introduces two factors which are used to adjust counts, one for the boarding number and the other for the alighting number. Similar to Cui's method [10], these factors are the ratios of the observed counts to their targets, which are the averages in this case. Multiplying boarding and alighting counts at each stop by these two factors respectively leads to the same total boarding and alighting counts.

Furth *et al.* propose another different solution [13] from the perspective of "passenger carry-over". If the total number of boarding doesn't equal to the total number of alighting, it is most likely that there are passengers onboard at the first stop or remaining at the last stop. It depends on which count is greater. Therefore, a pseudo stop is introduced to capture those carry-over passengers. In the case that the boarding count is greater than the alighting count, some passengers may want to get off in the next trip of the opposite direction. So a pseudo stop can be added in the downstream of the last stop and those carry-over passengers are assumed to alight at that virtual stop. For the contrary case where the alighting count is greater than the boarding count, there might be some passengers onboard from the

previous trip and those passengers can be assumed getting on at a pseudo stop which is added in the upstream of the first stop.

This pseudo stop method is also used to resolve the negative load problem [4] [14]. If a negative load problem is identified in APC data, adding one pseudo begin stop and one pseudo end stop helps to eliminate this issue. In details, through loads should be calculated by using the equation 2-1 in the section 2.1.2 for all stops at first, and then the maximum absolute negative load can be identified. This value should be considered as the number of passengers boarding at the pseudo stop in the beginning of the route and alighting at the pseudo stop at the end of the route.

2.2 Expressions of Transit OD Trip Estimation

The result of transit passenger OD trip estimations can be expressed in an OD flow matrix [4] [10], a probability flow matrix [3] [7] [15] [16] or an alighting probability matrix [11] [17]. The OD flow matrix represents travel flows between stop pairs. The probability flow matrix can be converted from normalization of the OD flow matrix. An entry in one probability flow matrix can be calculated by dividing the corresponding entry in the OD flow matrix by the total flow of the matrix and it is the probability that a random individual trip occur between two stops. The sum of all entries in a probability flow matrix equals to 1.

An alighting probability matrix represents probabilities of passengers from a certain stop alighting at its downstream stops. The alighting probability matrix

can be obtained by dividing each entry in the OD flow matrix by its row sum. Therefore the sum of each row equals to 1.

All of these three expressions are convertible, but several studies indicate a slight advantage in favor of the probability flow matrix and the alighting probability matrix. Sometimes researchers cannot obtain boarding and alighting data for all trips within their research time-of-day period, therefore to generate a route-level OD flow matrix requires some assumptions to scale up the OD flow matrix obtained from boarding and alighting data. This assumption may introduce another bias, while the probability flow matrix and the alighting probability matrix are tend to be fixed across transit trips and therefore reflects travel patterns better [11].

2.3 Route-Level Transit OD Trip Estimation Methods

This section reviews methods to conduct route-level transit OD trip estimations from APC boarding and alighting counts.

In a transit OD matrix, stops in rows are the origins of a trip and stops in columns are the destinations of a trip, and the number in each cell is the passenger flow, the probability flow or the alighting probability, from the origin stop to the destination stop in the different expressions. However, too many potential OD matrices can satisfy a given set of such sums. The task of transit OD trip estimation method is to find out the most appropriate one. [11]

2.3.1 Iterative Proportional Fitting Method

The Iterative Proportional Fitting (IPF) is a state-of-practice method used by agencies and institutions to estimate OD matrix from boarding and alighting counts [15]. It is also known as Bregman's balancing method [18] [19]. It has been proven that this method is practical in term of computing time. Its accuracy is acceptable as well from both empirical cases and simulation evaluations [3] [10].

Boarding and alighting counts for every stop along a bus route are required and referred to as marginal values. For one certain stop, its boarding count should be the sum of all passenger flows in the row of that stop, and its alighting count should be the sum of all passenger flows in the column of that stop.

Besides, a base matrix is required too. It is used as the start point of IPF calculation. The base matrix usually is an out-dated OD matrix, or estimations derived from other data sources. The result of IPF is proportional to the base matrix. In many practical cases, it is difficult to have a base matrix. Then, a null base matrix can be used. A null base matrix is a matrix with the same number in each cell. Usually the value is 1. Although a null base matrix performs well in some researches [7], the null base matrix may introduce some biases to the result [15] [20]. It implies that all stop-pairs have the same passenger flows, which is not realistic in most practical cases.

The IPF method is an iterative procedure. Numbers in each cell are multiplied by a factor in each iteration. The row factor is the ratio of observed

boarding count at each stop to the sum of estimated entries in the corresponding row from previous iteration, and the column factor is the ratio of observed alighting count at each stop to the sum of estimated entries in the corresponding column from the previous iteration. For the first iteration, estimated numbers come from the base matrix. By multiplying corresponding row factors, the sum of estimated boarding counts of each stop should equal to its observed boarding counts. However, the sum of estimated alighting counts usually do not equal to the observed alighting counts. Therefore, the estimated number in each cell should be multiplied by a column factor in order to achieve equivalence between the estimated alighting counts and the observed alighting counts. This is one round of this iterative process. In the end of this round, the difference of estimated boarding counts and observed boarding counts is checked again. If the difference is larger than a pre-defined convergence threshold, then the iterative process is continuing, until the difference of estimated counts and observed counts, no matter the boarding counts or the alighting counts, meets the convergence threshold. The convergence of the IPF procedure was proven by Fienberg in 1970 [21].

Given in the k iteration, the row factor for original stop i is a_i^k and the column factor for the destination stop j is b_j^k , the estimated passenger flow of the stop pair (i, j) in OD flow matrix can be formulated as:

$$\hat{t}_{ij}^k = a_i^k b_j^k \hat{t}_{ij}^{k-1} \quad (2\text{-}2)$$

, where

$\hat{t}_{ij}^{k}$ = the esitmated passenger flow between the stop i and stop j in the k iteration. [3] [4]

For the first iteration, the estimated passenger flow between stop i and stop j is the number of cell (i,j) in the base matrix, t_{ij}^{0}, therefore the equation 2-2 could be changed to

$$\hat{t}_{ij}^{k} = a_i b_j t_{ij}^{0} \quad (2\text{-}3)$$

, where

$a_i = \prod_{n=1}^{k} a_i^n$. This is the product of row factors in all iterations.

$b_j = \prod_{n=1}^{k} b_j^n$. This is the product of column factors in all iterations.

Ideally, the final estimated flow of every stop pairs should stratify the following constraints:

$$\sum_{j=i}^{N} \hat{t}_{ij} = P_i, i = 1, 2, 3, \dots, N \quad (2\text{-}4)$$

$$\sum_{i=1}^{j} \hat{t}_{ij} = Q_j, j = 1, 2, 3, \dots, N \quad (2\text{-}5)$$

, where

$\hat{t}_{ij}$ = the final estimated flow of the stop pair (i,j) in the OD flow matrix,

P_i = the observed boarding counts of the stop i,

Q_j =the observed alighting counts of the stop j,

N = the number of stops along the observed route.

The iterative process of IPF method is illustrated in Figure 1 as below,

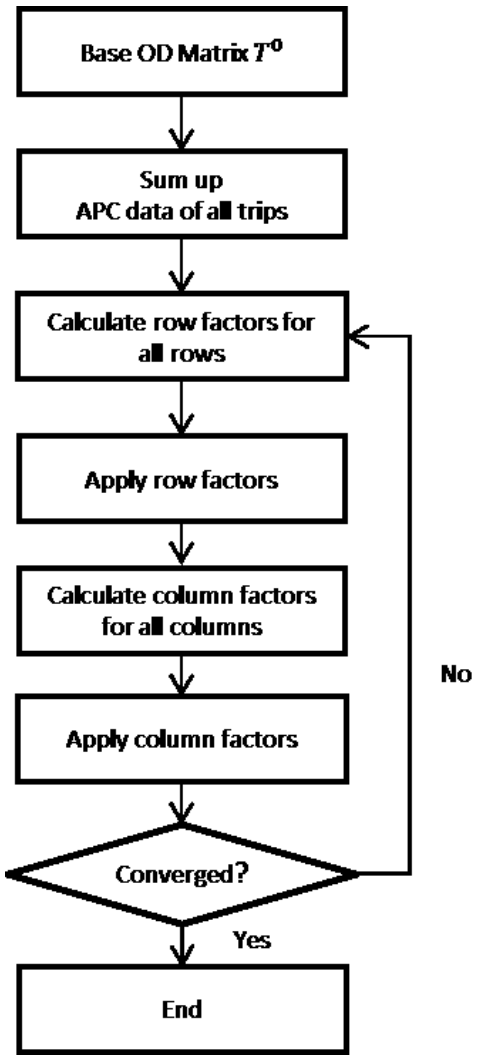


Figure 1 IPF Method Process

2.3.2 Iterative Proportional Fitting with an Iteratively Improved Base

Ji, Mishalani and McCord proposed an improvement to the IPF method [15]. Their method takes the advantage of large amount of APC data collected by transit agencies routinely to address the issue of missing a good priori base matrix.

Unlike the original IPF method, this method uses a base matrix which evolves iteratively after each round of computing. In the first iteration, it applies the original IPF method with an arbitrary base matrix to each bus trip. The outputs are trip-level OD flow matrices of all trips. All of those trip-level OD flow matrices will be aggregated into a route-level OD flow matrix. This aggregated OD flow matrix will be converted to a probability flow matrix which will be used as the base matrix for the next iteration. Several convergence thresholds can be checked to decide whether or not to continue this process, but the authors suggested using the difference of absolute value of each cell of probability flow matrices from two sequential rounds from the practical perspective. When this change is less than a pre-defined threshold, then the whole process shall finish.

The Figure 2 below depicts the process of this iterative proportional fitting with an iteratively improved base (IPF-IB) method.

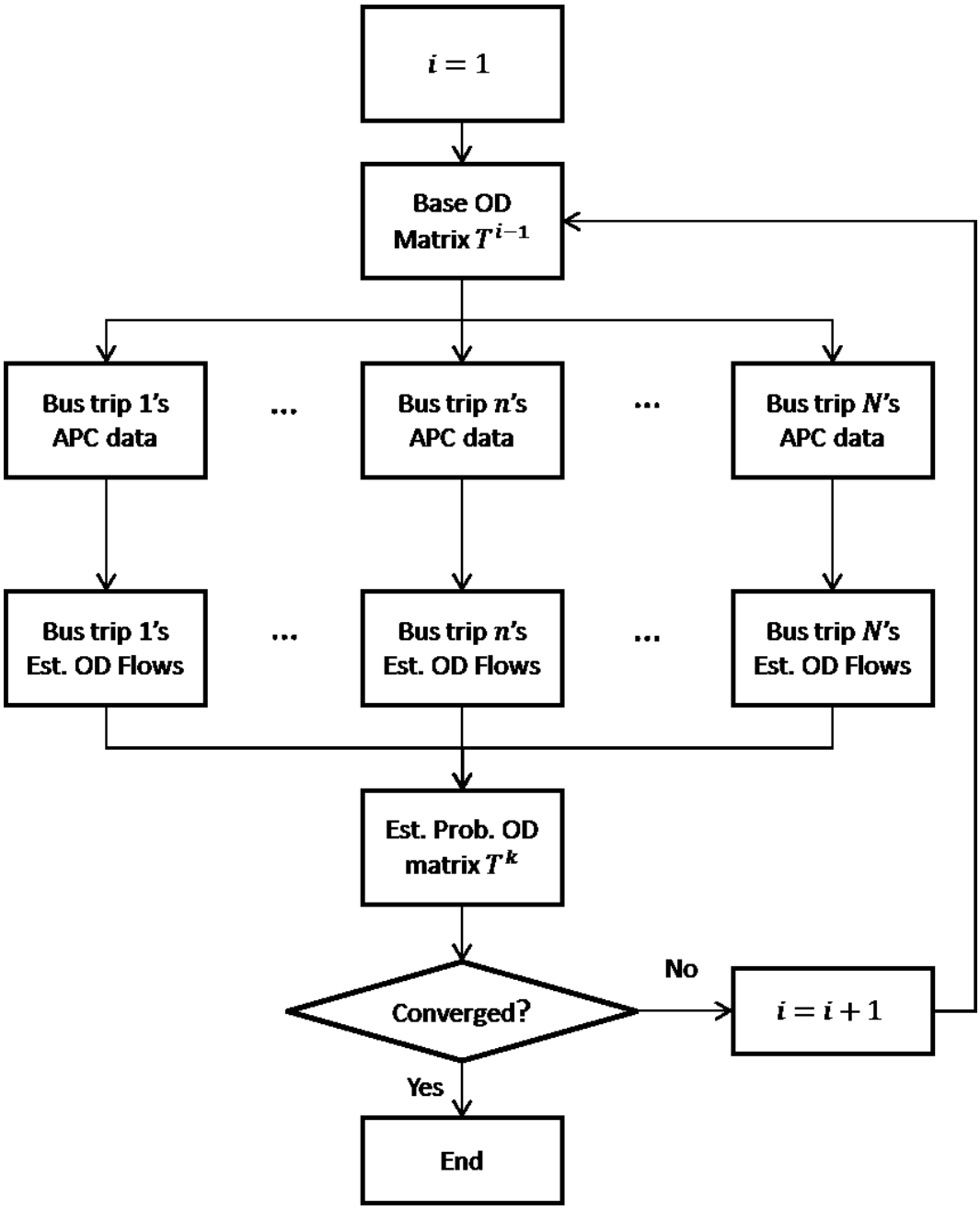


Figure 2 IPF-IB Method Process [15]

First of all, the iteration counter i is initialized as 1 and a null base matrix is selected as T^0, although the authors mentioned that any arbitrary matrix could be chosen [15].

The second step is to apply the original IPF method introduced in section 2.2.1 to each trip n individually, $n = 1, 2, 3, \dots, N$. The inputs of the IPF method are APC boarding and alighting counts of that trip and the base matrix T^{i-1} is generated from the previous iteration. The outputs of this step are N trip-level OD flow matrices of all N trips respectively.

The third step is sum up all N trip-level OD flow matrices to get one route-level OD flow matrix, and then normalize this OD flow matrix to a probability flow matrix T^i by dividing each entry in the route-level OD flow matrix by the total passenger volume. Since the trip-level OD flow matrices are calculated separately in each iteration, the aggregated route-level OD flow matrix is different from iteration to iteration. This output is observed to be converged after many iterations in the empirical study, although this phenomenon has not been proven theoretically [15].

The last step of one iteration is to check if the output of the third step, the probability flow matrix, is converged or not, or if the pre-specified number of iterations have been executed. If yes, then the whole method will stop; or this method will continue from the second step described above.

Compared to the original IPF method, the IPF-IB method uses an iterative process to improve the base matrix, which comes from the immediate preceding iteration. Since the base matrix is supposed to be better and better after iterations, the IPF-IB method overcomes the limitations imposed by the null base matrix at the end [15] .

2.3.3 Li and Cassidy Method

Li and Cassidy developed a route-level transit OD trip estimation algorithm [11] based on the Tsygalnitsky method [22] which uses boarding and alighting counts but does not require a base matrix.

This algorithm splits stops into two groups: major stops and minor stops. Generally stops which have high passenger volumes are usually considered as major stops and others are minor stops. Authors assume the alighting probabilities are different at a certain stop for passengers who board the bus at major stops and minor stops. An auxiliary parameter α represents the relationship between them:

$$\frac{p_{as}}{p_{bs}} = \frac{1-\alpha}{\alpha} \tag{2-6}$$

, where

p_{as} = the alighting probability at stop *s* for passengers who get on the bus at a major stop,

p_{bs} = the alighting probability at stop *s* for passengers who get on the bus at a minor stop; and

$\alpha \in (0, 1)$.

By applying the Bayesian analysis, the estimated number of passengers who alight at stop *s* from major stops can be calculated as:

$$\hat{n}_a = \frac{(1-\alpha)N_a}{(1-\alpha)N_a+\alpha N_b} n \tag{2-7}$$

, where

$\hat{n}_a$ = the estimated number of passengers who get off the bus at this stop *s* from major stops,

N_a = the sum of onboard passengers who get on the bus at major stops in the upstream of this stop *s*,

N_b = the sum of onboard passengers who get on the bus at minor stops in the upstream of this stop *s*,

n = the number of passengers who alight at this stop *s*,

α = the auxiliary parameter for this stop *s*.

$\hat{n}_a$ can be split to upstream major stops proportionally according to the ratio of their boarding counts to N_a. The estimated number of passengers from minor stops $\hat{n}_b = n - \hat{n}_a$ and it is assigned to upstream minor stops in the same way as $\hat{n}_a$. As a result, numbers of passengers from each upstream stop to the current stop are estimated and an entire OD flow matrix can be obtained by applying this calculation to all stops along this route.

Theoretically, the equation 2-7 may yield $\hat{n}_a > N_a$ when $\frac{\alpha}{1-\alpha} < \frac{n-N_a}{N_b}$. In this case, $\hat{n}_a = N_a$. And in the case that $\hat{n}_b > N_b$, $\hat{n}_a = n - N_b$.

All major stops are supposed to have a constant value, α_a as their auxiliary parameter and so do minor stops, which is usually different from α_a and represented as α_b.

Besides, it also assumes there is a minimum riding distance for passengers. Passengers who travel less than this distance may prefer to be onboard, while those who have travelled greater than it have priority in alighting.

In the paper [11], the researchers tested a range of minimum riding distances which vary from 0 km to 4.8 km in increments of 0.4 km. 0.4 km is usually considered as the maximum walking distance [9]. The case study conducted in their research produced the optimal estimation when the minimum riding distance is 3.2 km [11].

The comprehensive process is described as below:

Step 1: Estimate OD flow matrices for all trips for given α_a and α_b according to the equation 2-7 and APC data of each stop.

Step 2: Aggregate all OD flow matrices and then convert it to an alighting probability matrix.

Step 3: Assess the fitness of the alighting probability matrix for given α_a and α_b by checking differences of the estimated alighting counts based on this alighting probability matrix and the observed alighting counts for all trips.

Step 4: Repeat step 1 to step 3 for different α_a and α_b. The alighting probability matrix which fits trips the best is the result of this method.

In Li and Cassidy's study [11], they began from $\alpha_a = 0.1$ and $\alpha_b = 0.1$, increased these two auxiliary parameters with 0.1 step by step until $\alpha_a = 0.9$ and

$\alpha_b = 0.9$ for a real bus route served by AC Transit in the San Francisco Bay Area. $\alpha_a = 0.1$ and $\alpha_b = 0.3$ are the best choice, which means the alighting probability at major stops for passengers from major stops is about 9 times than passengers from minor stops, while at minor stops, passengers from major stops are about twice as likely to alight as those from minor stops.

2.3.4 Markov Model Method

Baibing Li investigated statistical inference in estimating a public transit OD matrix with APC data [17]. The author assumes that the alighting probability at a stop depends on the state of passengers at the previous one stop and uses first order Markov model to characterize transition probabilities, which can be defined in the equation below:

$$Pr\{\xi_i = k|\xi_{i-1} = m\} = \begin{cases} q_i & if\ k = 0\ and\ m = 1 \\ 1 - q_i & if\ k = m = 1 \end{cases} \text{ for } i = 2, \dots, N-1$$

(2-8)

, where

$\xi_i =$ the passenger state at stop *i*, where $\xi_i = 1$ if a passenger is onboard and $\xi_i = 0$ if not.

$q_i =$ the alighting probability for a passenger at stop i when he or she is onboard at stop $i - 1$

$N =$ the number of stops.

The alighting probability for passengers who board at stop i and alight at stop j could be:

$$p_{ij} = Pr\{alighting\ at\ stop\ j\ |boarding\ at\ stop\ i\} =$$

$$Pr\{\xi_j = 0, \xi_{j-1} = 1, \dots, \xi_{i+1} = 1 | \xi_i = 1\} \quad (2\text{-}9)$$

, where

$$i = 1, \dots, N-1,$$

$$j = i+1, \dots N$$

Therefore,

$$p_{ij} = \begin{cases} q_j, & \text{if } j = i+1 \\ q_j \prod_{k=i+1}^{j-1}(1-q_k), & \text{if } j = i+2, \dots, N \end{cases} \quad (2\text{-}10)$$

The equation when $j > i+1$ could be derived according to the properties of Markov chains as follows:

$$p_{ij} = Pr\{\xi_j = 0, \xi_{j-1} = 1, \dots, \xi_{i+1} = 1 | \xi_i = 1\}$$

$$= Pr\{\xi_j = 0 | \xi_{j-1} = 1, \dots, \xi_i = 1\} Pr\{\xi_{j-1} = 1 | \xi_{j-2} = 1, \dots, \xi_i = 1\} \dots Pr\{\xi_{i+1} = 1 | \xi_i = 1\}$$

$$= q_j(1-q_{j-1}) \dots (1-q_i) = q_j \prod_{k=i+1}^{j-1}(1-q_k) \quad (2\text{-}11)$$

Then the author carries out the Bayesian analysis to estimate q_j with the assumptions that q_j follows a beta distribution with parameters α_j and β_j. An

estimation of q_j can be yielded by calculating the mean of the posterior distribution of this Bayesian analysis:

$$\hat{q}_j = \frac{(\alpha_j + Q_j)}{\alpha_j + \beta_j + \sum_{k=1}^{j-1}(P_k - Q_k)}, \quad j = 2, \ldots, N-1 \tag{2-12}$$

, where

$\hat{q}_j =$ the estimated alighting probability for a passenger at stop i when he or she is onboard at stop $i-1$,

$P_i =$ the observed boarding counts at the stop i,

$Q_j =$the observed alighting counts at the stop j.

$q_N \equiv 1$ because all passengers are supposed to alight at the last stop.

The hyper-parameters α_j and β_j of the beta distribution are usually unknown beforehand. The equation 2-12 can take non-informative prior, $\alpha_j = 1$ and $\beta_j = 1$ to estimate $\hat{q}_j$ for several trips from a sample. Then those estimates are considered as prior information to infer the hyper-parameters α_j and β_j. Once the hyper-parameters are determined, Markov transit probabilities can be estimated by using the equation 2-12, and then the alighting probability matrix can be calculated by substituting the equation 2-12 into the equation 2-10. This alighting probability matrix is the output of this method, and it is easy to reconstruct to an OD flow matrix by multiplying boarding counts at corresponding stops.

2.4 Bus Stops Grouping Method

McCord, *et al*, figured out that the size of stop-to-stop OD matrices may introduce difficulties in revealing critical flow patterns when the number of stops is very large. They proposed a method to group bus stops to reduce the size of OD matrices [23]. As a result, passenger flows between far fewer stop group pairs are estimated and represented, but the result is believed to capture important OD flow characteristics.

This method sums up entries of consecutive stops in the original OD matrix, which usually is a probability flow matrix, according to different group configurations to form a new OD matrix which has fewer cells, and then checks the similarity between this aggregated OD matrix and the original one. The group configuration which generates the most similar aggregated OD matrix is selected as the output of this grouping method.

The researchers suggested using squared Hellinger distance, HD^2, as the similarity measure of two OD matrices [24]. In order to make the aggregated OD matrix and the original one have the same dimensions, the aggregated OD matrix shall be disaggregated to a stop-to-stop OD matrix again by dividing entries of the aggregated OD matrix evenly into all feasible stop pairs that were previously aggregated into that entry. The squared Hellinger distance is defined as

$$HD^2 = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\sqrt{p_{ij}^o} - \sqrt{p_{ij}^d} \right)^2 \quad (2\text{-}13)$$

, where

p_{ij}^{o} = the probability flow of the OD pair of stop i and stop j in the original OD matrix.

p_{ij}^{d} = the probability flow of the OD pair of stop i and stop j in the disaggregated OD matrix.

N = the number of stops.

The researchers also noticed that the amount of possible group configurations may be very huge when there are many stops along a route. If a route has N stops and the M-group configuration, which divides all stops into M groups, is desired, the number of potential group configurations is

$$(N-1)!/((M-1)!\,(N-M)!) \tag{2-14}$$

If there are 80 stops along a route, the number of 8-group configurations is 2,898,753,715 and the number of 10-group configurations is 205,811,513,765. It is impossible to find out the best group configuration from such huge amount of configurations within acceptable computing time.

Therefore, a heuristic method is proposed to find an approximate result without consuming too much time. The method does not check every possible configuration for M groups but begins at 1 group or N groups, both of which only have one configuration, and proceeds iteratively to M groups.

The method beginning from 1-group configuration is named as "Top-Down" algorithm. In the m-th step, m groups will be determined. The way to

obtain $(m+1)$ groups is to split any group in the m-group configuration into two in the $(m+1)$-th step. $N-m$ new group configurations will be generated. The configuration which produces the most similar disaggregated OD matrix to the origin one will be selected as the $(m+1)$-group configuration. The process is repeating until the desired group number is reached.

The method beginning from the N-group configuration follows the similar logic but merges any two groups into one at each step instead of splitting one into two. This way is called "Bottom-Up" algorithm.

2.5 Summary of Literature Review

In this chapter, APC data collection systems and methods used to conduct route-level transit passenger OD trip estimation from APC data are reviewed.

There are several different types of APC data collection systems in the world and the followings are widely used ones: Infrared Light Beams, Pressure Sensitive Mat, Passive Infrared Sensors and Cameras. Their strengths and drawbacks are discussed in this chapter as well.

Before conducting transit OD trip estimations, APC data should be examined to make sure its reliability [25]. APC data may have the imbalance issue and the negative load problem. They may be caused by the collection system itself or passenger behaviours like "carry-over". Data preprocessing methods should be developed to address these two issues. Otherwise, these issues may break OD estimation methods [4].

Four methods which can estimate route-level transit OD trips from APC data are reviewed: the IPF method, the IPF-IB method, the Li and Cassidy method, and the Markov model method.

The IPF and IPF-IB method require a base matrix and use an iterative process to approach their results. The IPF-IB can improve the base matrix iteratively so that it can relieve biases caused by a null base matrix.

The Li and Cassidy method develops a model to calculate alighting counts for each entry in the OD flow matrix without the base matrix. This method assumes the alighting probabilities at a certain stop is different for passengers from major stops and minor stops, and proposes an auxiliary parameter α to represent their relationship. It produces estimations by testing different α which vary from 0.1 to 0.9 and chooses the best-fit estimation as the result.

The Markov model method develops a closed-form equation to calculate alighting probabilities based on a first order Markov model. The OD flow matrix can be reconstructed based on the alighting probability matrix estimated by this method.

The bus stop grouping method is used to group consecutive stops based on a probability flow matrix. It can help to reduce the size of the estimated stop-to-stop matrix and reveals important flow patterns.

CHAPTER 3. CASE STUDY AND METHOD IMPLEMENTATION

This chapter introduces the Route 1 of the Edmonton Transit System (ETS) and the APC data collected by ETS. Implementation of methods to preprocess APC data, conduct transit passenger OD trip estimations from APC data and group stops are also covered by this chapter.

3.1 Introduction of Edmonton Transit System

The Edmonton Transit System (ETS) is the public transit agency which is totally owned and operated by the City of Edmonton. It provides regular buses services, Light Rail Train (LRT) services and door-to-door Disabled Adult Transit Service (DATS) for people living in the whole city area and the Capital Region surrounding the city.

ETS is the 6th largest transit agency all around Canada. It has over 2,200 employees, about 1,000 buses, 74 Light Rail Vehicles (LRV), 98 DATS vehicles, 6 Transit Garages, 25 Transit Centers, over 6803 stops, 15 LRT stations and 20.3 KM of LRT track. It provides more than 2 million service hours annually and delivers 80.2 million rides, including bus and LRT. [26]

The Figure 3 shows the system map of the Edmonton Transit System for Day Services.

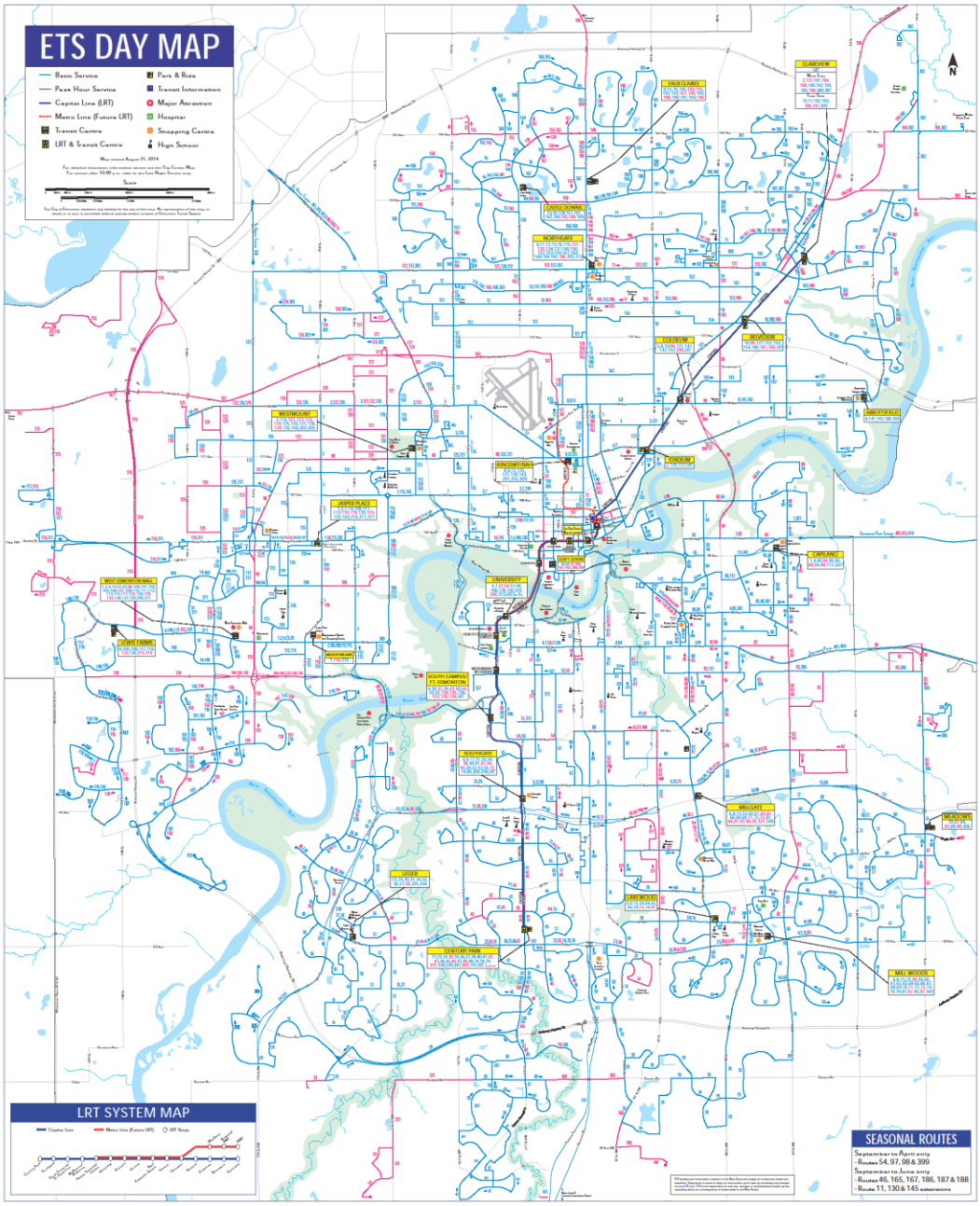


Figure 3 ETS Map - Day Service [27]

3.2 Introduction of Route 1

The Route 1 is one of the most important routes in ETS. It runs between the Capilano Transit Center and the West Edmonton Mall Transit Center. The Figure 4 illustrates the map of the Route 1.

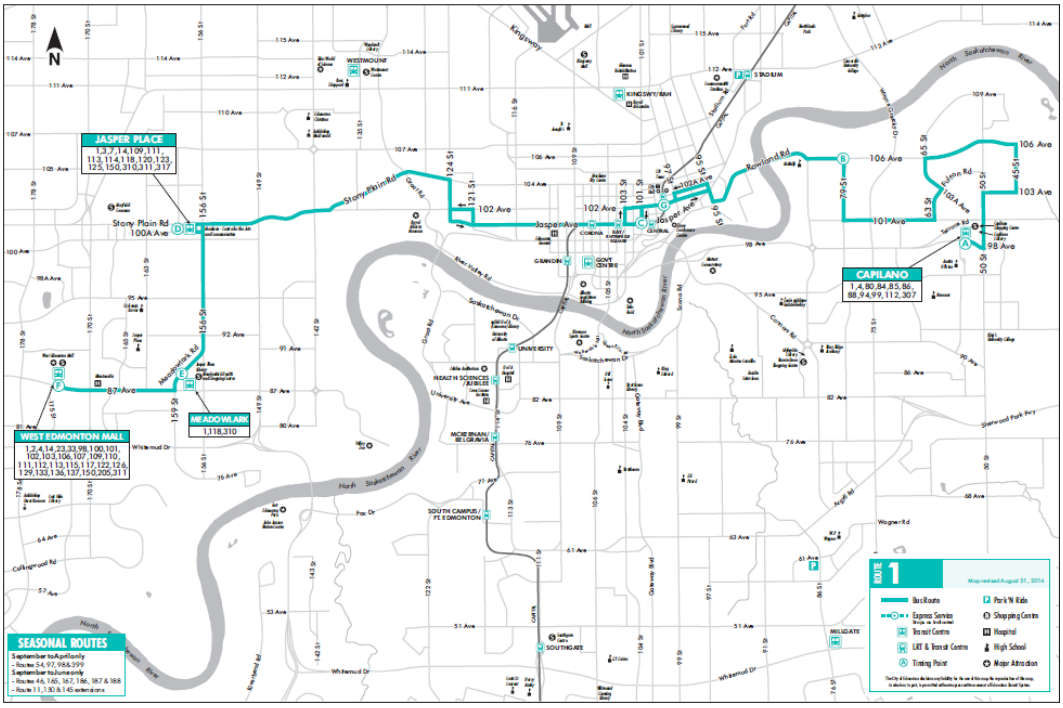


Figure 4 The Map of the Route 1 [28]

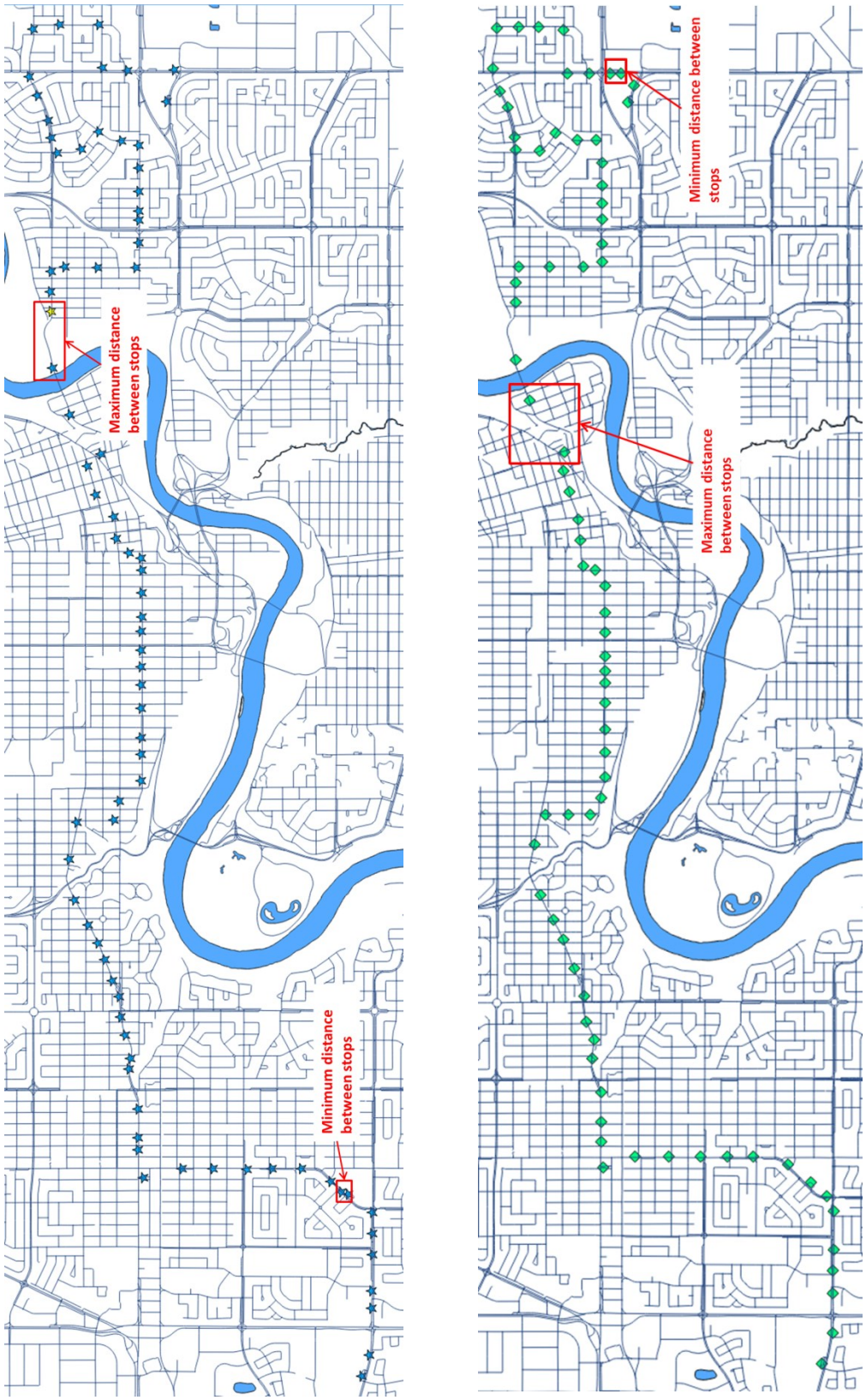
The total length of this route is about 22.5 KM. There are 80 stops along the whole route. 4 stops among them are in the Capilano Transit Center, the Jasper Place Transit Center, the Meadowlark Transit Center and the West Edmonton Mall Transit Center respectively.

The average distance between stops in the westbound direction from the Capilano Transit Center to the West Edmonton Mall Transit Center is 270 meters; the maximum distance in this direction is 586 meters between the stop 2547 and the stop 2276, which are located in 83 Street and 106 Avenue, and Dawson Bridge and Rowland Road respectively. There is a big golf club nearby between these two stops and may not have high demand for transit stops; the minimum distance is about 71 meters between the stop 5580 and the stop 5301, which are

located in Meadowlark Road and 89 Avenue, and Meadowlark Transit Center respectively.

The average distance in the eastbound direction between stops is 267 meters; the maximum distance is 648 meters between the stop 1188 and the stop 1173, which are located in 95 Street and Jasper Avenue, and 98 Street and Rowland Road respectively; the minimum distance is 120 meters between the stop 2192 and the stop 2291, which are located in 50 Street and 101 Avenue, and 50 Street and 98 Avenue respectively.

The Figure 5 depicts stops along the Route 1 in both directions and highlights the maximum and minimum distances between stops.



(a) stops in WB direction (b) stops in EB direction

Figure 5 The Route 1 Stops

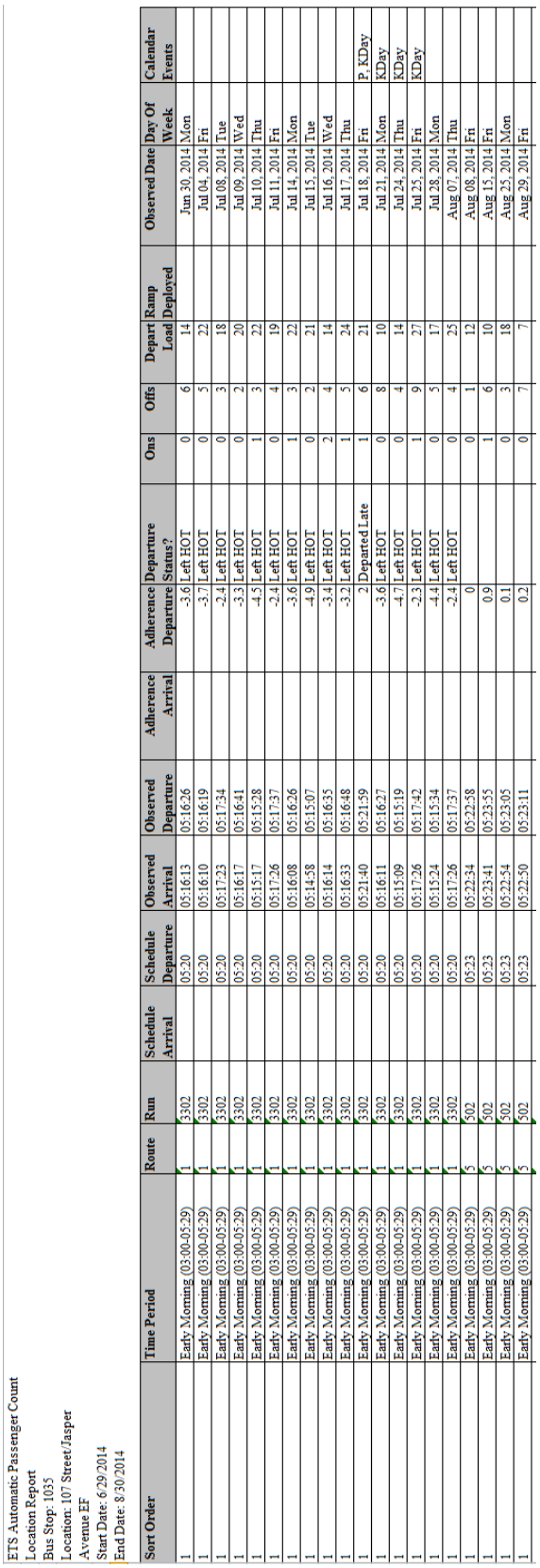
3.3 APC Data

3.3.1 APC Data in ETS

ETS doesn't install the APC data collection system on all buses. There are about 200 buses equipped with this system. They are assigned to different routes from day to day in order to cover the whole transit system as much as possible. This circulation through all routes allows ETS have a good understanding about how the whole transit network works.

ETS administrates databases of the APC data and provides some performance reports periodically, for example, the high load numbers, the low load numbers and significant delays. Among all of those reports, boarding and alighting counts of stops can be used to estimate OD trips for a route. The Figure 6 illustrates a sample of APC boarding and alighting counts.

The APC data files provided by ETS are not like the data in other researches which boarding and alighting counts at all stops along a route are given directly [4] [10] [15] [23]. Those files are grouped by stops. One file contains APC data of all routes which share that stop during a "signup", which is a term used in the APC data to indicate a report period that usually is two months. No counts are recorded by the system at stops where no passenger gets on or gets off. So in the extreme cases, some stops may only have few APC data records during a signup.



ETS Automatic Passenger Count
Location Report
Bus Stop: 1035
Location: 107 Street/Jasper Avenue EF
Start Date: 6/29/2014
End Date: 8/30/2014

Sort Order	Time Period	Route	Run	Schedule Arrival	Schedule Departure	Observed Arrival	Observed Departure	Adherence Arrival	Adherence Departure	Departure Status?	Ons	Offs	Depart Load	Ramp Deployed	Observed Date	Day Of Week	Calendar Events
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:13	05:16:26		-3.6	Left HOT	0	6	14		Jun 30, 2014	Mon	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:10	05:16:19		-3.7	Left HOT	0	5	22		Jul 04, 2014	Fri	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:17:23	05:17:34		-2.4	Left HOT	0	3	18		Jul 08, 2014	Tue	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:17	05:16:41		-3.3	Left HOT	0	2	20		Jul 09, 2014	Wed	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:15:17	05:15:28		-4.5	Left HOT	1	3	22		Jul 10, 2014	Thu	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:17:26	05:17:37		-2.4	Left HOT	0	4	19		Jul 11, 2014	Fri	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:08	05:16:26		-3.6	Left HOT	1	3	22		Jul 14, 2014	Mon	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:14:58	05:15:07		-4.9	Left HOT	0	2	21		Jul 15, 2014	Tue	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:14	05:16:35		-3.4	Left HOT	2	4	14		Jul 16, 2014	Wed	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:33	05:16:48		-3.2	Left HOT	1	5	24		Jul 17, 2014	Thu	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:21:40	05:21:59		2	Departed Late	1	6	21		Jul 18, 2014	Fri	P, KDay
1	Early Morning (03:00-05:29)	1	5302		05:20	05:16:11	05:16:27		-3.6	Left HOT	0	8	10		Jul 21, 2014	Mon	KDay
1	Early Morning (03:00-05:29)	1	5302		05:20	05:15:09	05:15:19		-4.7	Left HOT	0	4	14		Jul 24, 2014	Thu	KDay
1	Early Morning (03:00-05:29)	1	5302		05:20	05:17:26	05:17:42		-2.3	Left HOT	1	9	27		Jul 25, 2014	Fri	KDay
1	Early Morning (03:00-05:29)	1	5302		05:20	05:15:24	05:15:34		-4.4	Left HOT	0	5	17		Jul 28, 2014	Mon	
1	Early Morning (03:00-05:29)	1	5302		05:20	05:17:26	05:17:37		-2.4	Left HOT	0	4	25		Aug 07, 2014	Thu	
1	Early Morning (03:00-05:29)	5	502		05:23	05:22:34	05:22:58		0		0	1	12		Aug 08, 2014	Fri	
1	Early Morning (03:00-05:29)	5	502		05:23	05:23:41	05:23:55		0.9		1	6	10		Aug 15, 2014	Fri	
1	Early Morning (03:00-05:29)	5	502		05:23	05:22:54	05:23:05		0.1		0	3	18		Aug 25, 2014	Mon	
1	Early Morning (03:00-05:29)	5	502		05:23	05:23:50	05:23:11		0.2		0	7	7		Aug 29, 2014	Fri	

Figure 6 A Sample of APC Data at the Stop 1035 in the Route 1

A stop's APC data contains many items. Those required conducting transit OD trip estimation for a route from APC data are listed below:

- Time Period

  Time period indicates when a bus is operating. This item is used to classify APC data into different time groups.

- Route

  This item is used to identify the route that this research is interested in.

- Run

  Run means a specific bus on a route within one day. The same run number may be shared between different routes in different days, but it is unique on one route within the same day.

- Observed Departure

  This item indicates when a bus leaves a stop.

- Ons

  This item counts how many passengers get on at this stop.

- Offs

  This item counts how many passengers get off at this stop.

- Observed Date

  This item figures out when the data record is collected. At some stops, their observed date is not continuous during a signup because no boarding and alighting passengers at those stops or no APC-equipped buses operating on those missing dates.

3.3.2 APC Data of Route 1

The APC data set used in this research has about 52,500 APC data records which are collected on the Route 1 during the signup from June 30, 2014 to August 29, 2014. Those data records cover 45 days within this signup in total. There is no bus mounted with the APC system serving as the Route 1 on those missing days because APC buses are shifted among the all ETS fleets in order to maximize the coverage of APC data. Moreover, the numbers of runs vary every day, which means ETS assigns different numbers of APC buses to operate on the Route 1 every day. Maximum of 10 runs is observed on July 25, and minimum of 1 run occurs on July 1. On average, there are 4 runs per day during the signup.

3.4 Data Preprocessing

3.4.1 Extract Trips from APC Data

Formats of APC Data stored in transit agencies may be different for reasons. Some agencies have high level APC data which contains boarding and alighting counts, locations, trips and so on, while others may store low level APC data directly, like a record of each sensor beam being broken [14]. However, all transit OD trip estimation methods need to deal with a trip's boarding and alighting counts at stops along a route.

This research defines a data structure for a trip which contains all necessary information for transit OD trip estimation methods. A trip is defined as a bus running once from the first stop to the last stop along the route. Its data

structure should contain the following items so that transit OD estimation methods can process it:

Table 2 Trip Data Structure

Item	Description
Route	(*Optional*) This is not necessary for route-level transit OD estimation, but should be required for transit network OD estimation.
Date	The date when this trip occurs.
Time-of-day	The time-of-day period that this trip occurs.
Direction	The direction of this trip.
Ons Sequence	The sequence of raw boarding counts in the order of stops from the first stop.
Offs Sequence	The sequence of raw alighting counts in the order of stops from the first stop.
Adjusted Ons Sequence	The sequence of adjusted boarding counts after resolving the imbalance issue and the negative load problem.
Adjusted Offs Sequence	The sequence of adjusted alighting counts after resolving the imbalance issue and the negative load problem.

According to the features of the APC data provided by ETS, the following algorithm is proposed to extract trips:

Step 1: Choose the APC data collected for the target route at all stops along that route.

Step 2: Pick a date from the signup.

Step 3: Classify the APC data into groups by different run values. There are several runs on one route per day. Data which has the same run value is classified into the same group.

Step 4: Select one group and sort all APC data within this group according to the ascending order of "Observed Departure" time. This time stamp indicates the actual time when a bus leaves a stop. It helps to sort all stops in time sequence. This sequence should follow the same stop order in the schedule of this route.

Step 5: Split stops in the sequence into trips. If one stop is the first stop in the sequence, it is definitely the start of a new trip, no matter if it is the first stop of the route designated by the schedule or not. If a stop matches the last stop of this route, then this stop is the last stop of this trip and all stops between the first stop of this trip found previously and this stop belong to this trip. Once a trip is identified, the last stop of this trip is treated as the first stop of the next trip. This step repeats until all stops in the sequence have been checked.

Step 6: Repeat Step 4 and Step 5 until all run groups in the picked date have been checked.

Step 7: Repeat Step 2 to Step 6 until all dates in the signup have been checked. At the end of this step, all APC data are converted to trips.

The trip extraction process is illustrated in Figure 7.

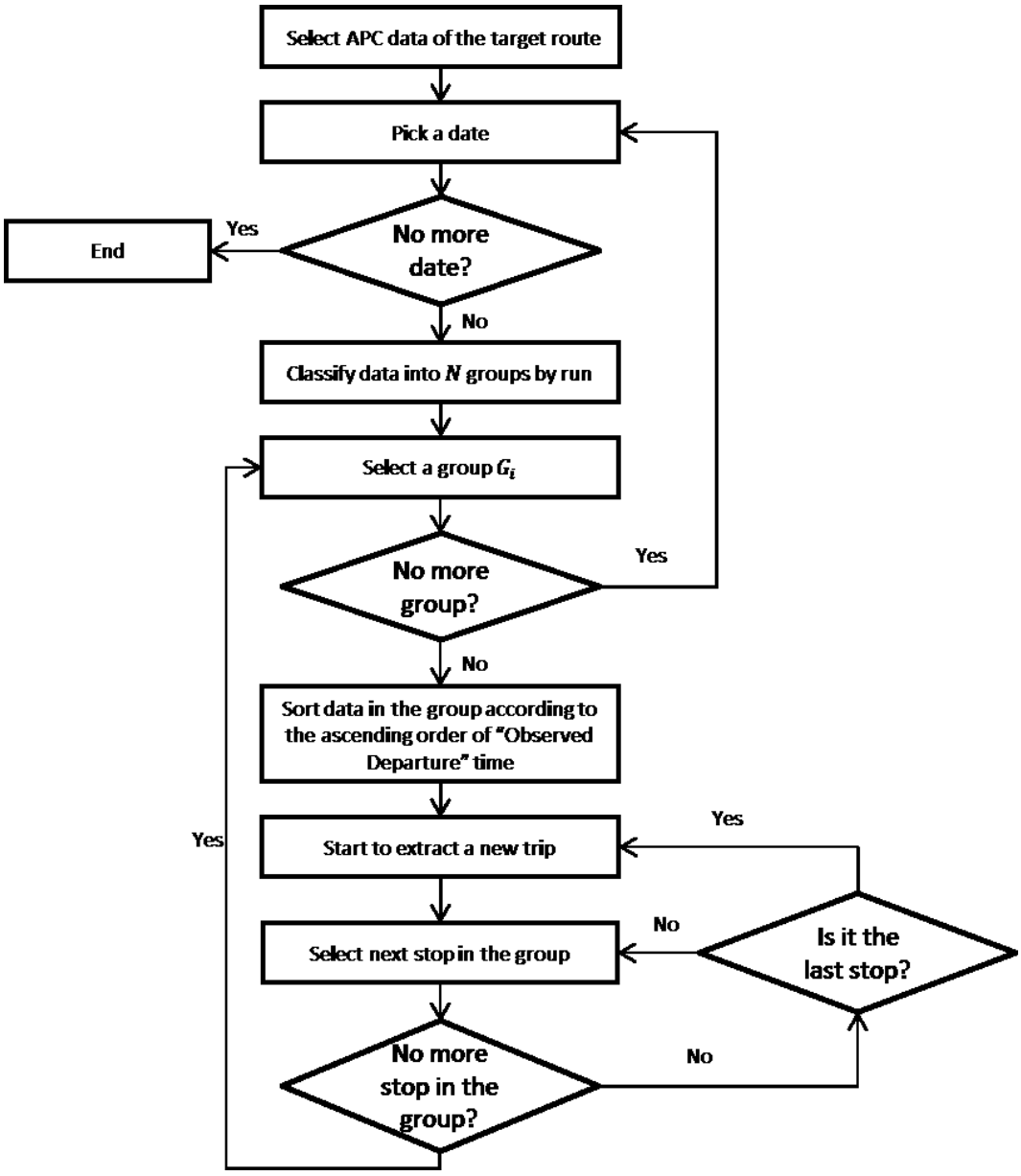


Figure 7 The Method of Extracting Trips from Stop-Based APC Data

One noteworthy thing is that not all stops in the schedule are included in one trip because if no passengers board or alight at a stop then the APC system

doesn't have data records for that stop. A simple validation step can be applied to check if the stop sequence of each trip follows the stop sequence in the schedule. This step just compares the sequence of the stops in one trip to the stop sequence in the routes' schedule and make sure there is no stop in wrong place in the sequence. The validation step performed in this research indicated the algorithm proposed in this section to extract trips from the stop-based APC data works as expected. Those missing stops in a trip are inserted into the trip's stop sequence with zero boarding and alighting counts.

3.4.2 Balance Trips' APC Data

As mentioned in the literature review section, APC data may contains errors because of system faults or passengers' "carry-over" behaviours. A balance procedure is required to meet the well-accepted assumptions in transit passenger OD trip estimations:

- No passenger alights at the first stop of a route.
- No passenger boards at the last stop of a route.
- The total number of boarding counts equals to the sum of alighting counts.

Besides, many ETS routes are operating in a loop. The last stop of the previous trip is the also the first stop of the next trip. The APC collection system doesn't split boarding and alighting counts into two different stops. Therefore, there are other assumptions in this research:

- The number of alighting counts at the first stop of a route is considered as the alighting counts at the last stop of the last trip from the opposite direction.

- The number of boarding counts at the last stop of a route is considered as the boarding counts at the first stop of the next trip.

In this research, the idea that matching the smaller APC counts, which might be boarding counts or alighting counts, to the larger APC counts is adopted. The steps of this algorithm are:

Step 1: Sum up boarding counts of all stops except the last stop.

Step 2: Sum up alighting counts of all stops except the first stop.

Step 3: Compute the ratio of the two sums and multiply this ratio to all stops' APC counts whose sum is smaller.

However, if difference between the total boarding counts and the total alighting counts of a trip is too large, there might be some significant errors in the APC data. In that case, researchers usually consider the data quality of that trip is not good and abandon that trip [4] [10]. In this research, if the percentage of the difference to the smaller APC counts of a trip is greater than 20%, that trip will be abandoned.

3.4.3 Resolve Trips' Negative Load

The negative load problem means the number of alighting passengers at one stop is greater than the number of onboard passengers approaching this stop. It may happen when some passengers get off immediately after they get on, or it is caused by APC system failures. This problem breaks estimation methods. For example, the IPF method cannot converge and results in oscillation in various accumulation points under this issue. Pukelsheim and Simeone proved this in [29]

by using L_1-error analysis. The way to calculate negative load in fact is a simplified case of the L_1-error function.

The method used in this research to deal with negative load problem is the same as introduced in the literature review section. A slight modification is that no pseudo stops are added before or after route terminals. The steps of this method are described in details as below:

Step 1: Calculate arrival load $L^n_{arrival}$ at stop n, which means how many passengers arrive at this stop.

$$L^n_{arrival} = \sum_1^{n-1} P_i - \sum_1^{n-1} Q_i \quad (3\text{-}1)$$

, where

$L^n_{arrival}$ = arrival load at stop n

P_i = boarding counts at stop i,

Q_i = alighting counts at stop i

Step 2: Calculate through load $L^n_{through}$ at stop n, which means how many passengers travel from upstream stops to downstream stops.

$$L^n_{through} = L^n_{arrival} - Q_n \quad (3\text{-}2)$$

, where

$L^n_{through}$ = through load at stop n

Q_n = alighting count at stop n

The number of passengers get off at this stop should be less than the number of arrival load of this stop. Otherwise, the negative load problem happens.

Equation 3-1 and equation 3-2 can be combined to get the following equation:

$$L^{n}_{through} = \sum_{1}^{n-1} P_i - \sum_{1}^{n} Q_i \quad (3\text{-}3)$$

Step 3: If no negative load problem occurs, this resolving process finished. Otherwise, go to step4.

Step 4: Select the negative through load which has the maximum absolute value and then add this absolute number to the boarding counts at the first stop of the trip. In order to keep the balance of boarding and alighting counts, this number should be added to the alighting counts at the last stop of this trip as well.

3.4.4 Homogenous Trip Groups

The travel patterns vary in different running directions and time-of-day periods. In this research, homogenous trip groups are defined according to these two factors in order to classify trips which share similar travel patterns. All transit OD trip estimation methods should be applied to those homogenous trip groups.

ETS has already grouped the Route 1's APC data into six time-of-day periods:

- Early Morning: 3:00 ~ 5:29
- AM Peak: 5:30 ~ 8:59
- Midday: 9:00 ~ 14:59
- PM Peak: 15:00 ~ 17:59
- Early Evening: 18:00 ~ 21:59
- Late Evening: 22:00 ~ 24:59

And the Route 1 has two running directions, the eastbound direction and the westbound. Therefore, there are 12 homogenous trip groups in this research.

3.5 Estimation Methods Implementation

The result of transit OD trip estimation in this research is expressed in both of the probability flow matrix and the alighting probability matrix. The probability flow matrix can be used to figure out high demand areas along the route and the alighting probability matrix is employed to assess fitness of estimations and analyze passengers' preference of alighting at some stops.

3.5.1 IPF Method

ETS did not conduct on-board surveys for the Route 1 or OD estimations before. There is no base matrix available for the IPF method. Therefore, the null base matrix is used as the base matrix for OD trip estimation in this research.

The IPF method introduced in the literature review section is applied to each trip in a homogenous trip group. The results are OD flow matrices of those trips.

In this research, the threshold of convergence is 0.000001, which ensures sums of rows and columns in OD flow matrices match boarding and alighting counts respectively as much as possible and the method can finish the computation in an acceptable time.

OD flow matrices of all trips are summed up to an aggregated OD flow matrix for that homogenous trip group and then the probability flow matrix and

the alighting probability matrix can be converted from that aggregated OD flow matrix.

3.5.2 IPF-IB Method

It is the same as the IPF method in this research that a null base matrix is selected as the initial base matrix in the first iteration. However, the IPF-IB method uses an iteratively improved base matrix in following iterations. In each iteration, the original IPF method is applied to all individual trips in a homogenous trip group.

A convergence threshold is defined to measure if the difference between the result, a probability flow matrix, of this iteration and the one from previous iteration is small enough. If yes, this probability flow matrix is the final result of this method. Otherwise, it is used as a base matrix for the original IPF method in next iteration.

The convergence threshold in this research is 0.000001, which allows having 1 person different in every 1,000,000 population between two different estimations. This condition makes sure the result has a high precision in this case study in the City of Edmonton, which has a population around 90,000 in 2014 [30]. In the research, all homogenous trip groups can archive the convergence threshold within 300 iterations.

3.5.3 Li and Cassidy Method

In this research, major stops selections have been found having a critical impaction on the result in this method. However, in the Li and Cassidy's paper, they didn't either analysis this impaction or figure out how to select major stops to optimize the result. They designated 10 stops along the route qualitatively according to their land use characteristics like easy access to train stations or shopping malls [11]. But it is not guaranteed to obtain the optimal OD matrix based on major stops selected by this means.

This research improves the original algorithm by adding extra iterations of testing different major stop groups which are determined by stops' APC counts. Major stops can be those stops which have high boarding counts, or high alighting counts, or both. Therefore, a major stop group can be determined by top n ($1 \leq n \leq 80$) stops according to their boarding counts, alighting counts, or sum of them. The improved method will assess estimated results for all major stop groups and choose the best-fit one as the final output of this method. In this research, the major stop group which consists of top 20 high-alighting-counts stops produces the best-fit estimation.

The minimum riding distance is another factor which impacts the final result and it may be vary in different routes according to time-of-day, geometry design, passenger trip purposes and so on. The range used in the paper with increments of 0.4 km [11] cannot find an available value other than 0 km for the

minimum riding distance in the Route 1 in ETS. Therefore, the increment of minimum riding distance is reduced to 0.1 km in this research.

3.5.4 Markov Model Method

In this research, the Markov transition probability is assumed to follow the same beta distribution $beta(\alpha_j, \beta_j)$ as in the paper [17]. One-third trips in a homogeneous group are selected randomly for estimating the hyper-parameters of the beta distribution. A non-informative prior, $\alpha_j = 1$ and $\beta_j = 1$, are used to estimate the Markov transition probabilities for those selected trips respectively by using the equation 2-12:

$$\{q_j^1, q_j^2, \dots, q_j^n\} \tag{3-4}$$

, where

$n =$ the number of selected trips,

$q_j^n =$ the Markov transit probability for stop j in the trip n.

Since the beta distribution with the upper bound $a = 1$ and the lower bound $b = 0$ has the following properties:

$$Sample\ Mean\ \bar{x} = \frac{\alpha}{\alpha+\beta} \tag{3-5}$$

$$Sample\ Variance\ S^2 = \frac{\alpha\beta}{[(\alpha+\beta)^2(\alpha+\beta+1)]} \tag{3-6}$$

The hyper-parameters α_j and β_j can be determined by the sample mean and sample variance of the set $\{q_j^1, q_j^2, \dots, q_j^n\}$. And then α_j and β_j are used as the

prior knowledge to estimate the Markov transition probability of the stop j for the entire homogenous trip group.

According to the equations 2-10 and 2-12, it is easy to obtain the alighting probability matrix for all trips in a homogenous trip group and then reconstruct the OD flow matrix by multiplying passenger boarding counts.

3.6 Overall Fitness Measure

This research uses an overall fitness measure [11] to test the fitness of estimated results. This overall fitness measure is calculated based on the vehicles' average load, which is calculated by dividing a vehicle's revenue passenger kilometers by the route distance.

$$x = \frac{\sum_{i=1}^{N-1} L_i * d_i}{D} \tag{3-7}$$

, where

$x =$ the average load,

$D =$ the distance of route,

$L_i =$ the number of passengers onboard from the stop i to stop $i + 1$,

$d_i =$ the distance between the stop i and stop $i + 1$.

In the equation 3-7, $\sum_{i=1}^{N-1} L_i * d_i$ is a vehicle's revenue passenger kilometers of a trip. And

$$L_i = \sum_{i=1}^{i} (P_i - Q_i) \tag{3-8}$$

, where

$P_i =$ the boarding counts at the stop i,

Q_i =the alighting counts at the stop i.

For a trip j in a homogenous trip group, an actual average load x^jcan be calculated with observed boarding and alighting counts; an estimated average load $\hat{x}^j$ can be obtained with observed boarding counts and estimated alighting counts which are calculated by multiplying the alighting probability matrix of this homogenous trip group by the observed boarding counts.

Then, the overall fitness measure of this alighting probability matrix is defined as:

$$F = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{x}^j - x^j)^2} \quad (3\text{-}9)$$

, where

F = the overall fitness measure

J = the number of trips in a homogenous trip group.

This overall fitness measure is used not only to assess the fitness of alighting probability matrices under different parameters in the Li and Cassidy method, but also to evaluate the four methods implemented in this chapter.

A smaller overall fitness measure value means that alighting probability matrix fits the homogenous trip group better. Therefore, in the Li and Cassidy method, the alighting probability matrix which has the lowest value of the overall fitness measure is selected as the final result. In the evaluation of those estimation methods, the method which produces the alighting probability matrix which has the lowest value is considered to be better than others.

3.7 Stop Grouping Method Implementation

This research employs the heuristic top-down stop grouping algorithm. The Route 1 has 80 stops in both directions. In order to group them, these stops are numbered from 1 to 80 according to the stop sequence in one direction.

In the step $M = 1$, this 1-group configuration considers all stops in one group, represented as

$$G(1) = \{80\} \quad (3\text{-}7)$$

Indexes in a group configuration represent the last stops of those groups. For example, the 1-group configuration has only one group and the index of the last stop of that group is 80.

The probability flow matrix of the 1-group configuration is in fact a 1 x 1 matrix [100%], therefore this grouping process starts from the step $M = 2$, a 2-group configuration is studied:

$$G(2) = \{g_x, 80\}, \quad (3\text{-}8)$$

, where

g_x = 1, 2, 3, …, 79. It is the index of the last stop in the first stop group.

The way to generate $(m + 1)$-group configuration is to insert one more stop index in the m-group configuration. For example, if

$$G(m) = \{g_1, g_2, \dots, g_{\mathrm{m}-1}, 80\} \quad (3\text{-}9)$$

, then the configuration in the step $M = m + 1$ is

$$G(m+1) = \{g_1, g_2, \dots, g_{n-1}, g_n, g_{n+1}, \dots, g_m, 80\} \quad (3\text{-}10)$$

g_n is the newly inserted stop index in the step $M = m + 1$, g_{n-1} and g_{n+1} are existing stop indexes in the m-group configuration, and $g_{n-1} < g_n < g_{n+1}$.

This process checks every $(m+1)$-group configuration with the method introduced in the literature review section and selects the one whose squared Hellinger distance metric, HD^2, is the smallest as the initial group configuration for the $(m+2)$-group configuration. This process repeats until the specified group number is reached and the configuration with the minimum HD^2 is the recommended group configuration.

3.8 Summary of Case Study and Method Implementation

Stop-based APC Data from the Route 1 in ETS will be processed in this research. Before conducting transit OD estimations from those data, they should be pre-processed. Trips will be extracted from those stop-based APC data and then grouped into homogenous trip groups. The imbalance APC data issue and the negative load problem shall be resolved by the data pre-processing methods.

In this research, four estimation methods are implemented and enhanced. The IPF method and the IPF-IB method use a null base matrix as the initial matrix because no outdated transit OD matrix or existing estimation from other data sources is available. The Li and Cassidy method is improved in this research. Instead of designating major stops according to stops' land use properties,

different major stop groups are tested according to their APC data in an extra iteration. This improvement of the Li and Cassidy method helps to find the best-fit estimation. The Markov Model method takes one-third trips in a homogenous trip group in this research as prior information to estimate the hyper-parameters of the beta distribution which Markov chain transit possibilities are assumed to follow. In addition, the overall fitness measure is employed to evaluate these four estimation methods in this research.

The stop grouping method is also implemented to reduce the size of stop-to-stop OD matrix. In this research, a top-down heuristic algorithm is applied in this method, which can achieve an appropriate grouping result within a short time.

CHAPTER 4. RESULTS AND DISCUSSIONS

This chapter presents the outputs of data preprocessing on the APC data of the Route 1, route-level transit OD matrices estimated by the IPF method, the IPF-IB method, the Li and Cassidy method and the Markov Model method, and the result of stop grouping based on those OD matrices, as well as analysis and discussions about those results.

4.1 Route 1 APC Data Preprocessing

4.1.1 Trips Extraction

The Route 1 has 80 stops in both directions. There are no stops shared by trips in different directions except two terminals of the Route 1, which are the stop 5009 in the West Edmonton Mall Transit Center and the stop 2301 in the Capilano Transit Center. Those trips running between the stop 5009 and the stop 2301 are defined as "complete trip" and others are considered as "incomplete trip". There two reasons to have incomplete trips:

1) According to the Route 1's schedule, some trips start or end at the stop in the Jasper Place Transit Center or some major stops in the Downtown area instead of those two terminals.

2) Some trips are lack of APC data at the two terminals because there is no passenger boarding or alighting or APC system got

some errors that time. Since it is unlikely no passengers get on or get off at those two terminals, most of those incomplete trips should be caused by APC system faults.

Totally 1652 trips are extracted from all APC data. 1375 trips out of them are complete and other 277 trips are incomplete. This research is only interested in those 1375 complete trips.

4.1.2 Descriptive Analysis

All eligible trips can be grouped into 12 homogenous trip groups according to the running direction and the time-of-day period. Table 3 below lists the descriptive statistical result after applying the APC counts balancing and negative load resolving method to all homogenous trip groups.

Table 3 Descriptive Statistical Result of Eligible Trips

Time Period	Direction	Average Stops	Number of Trips	Passenger Counts	Average Load/Trip
Early Morning	W to C	-	0	0	-
	C to W	37.6	12	615	51.25
AM Peak	W to C	33.5	55	2500.26	45.46
	C to W	37	95	4626.69	48.70
Midday	W to C	35.1	141	8356.42	59.27

Time Period	Direction	Average Stops	Number of Trips	Passenger Counts	Average Load/Trip
	C to W	35.8	152	9090.17	59.80
PM Peak	W to C	41.2	79	6619.34	83.79
	C to W	40.2	57	4332.57	76.01
Early Evening	W to C	34.3	93	5552.5	59.70
	C to W	29.6	51	2228.86	43.70
Late Evening	W to C	34.5	10	598	59.80
	C to W	31.4	8	357.69	44.71

Although there are 80 stops along the Route 1, a trip has about 30 stops on average, but trips during PM Peak have about 10 more stops than other time periods.

In the PM Peak, the average load per trip is about 80 passengers, while the average load per trip in other time-of-day periods is around 52 passengers, which is the designed capacity of a bus in Edmonton. This column suggests that the Route 1 operation meets the demand in most time of a day except the PM Peak period. Considering this column and the “Average Stops” column together, the PM Peak is the most busy time-of-day period for the Route 1and the demand excesses the service supply of the Route 1 in that time period.

4.2 Route-Level Transit Passenger OD Trip Estimation

This thesis uses APC data collected in the direction from the Capilano Transit Center to the West Edmonton Mall Transit Center during the Midday period (9:00 ~ 14:59) to evaluate the four estimation methods: the IPF method, the IPF-IB Method, the Li and Cassidy Method and the Markov Model Method.

4.2.1 Estimation Results

The route-level transit passenger OD trip estimations in this research are presented in the form of the probability flow matrix, where an entry expresses the probability of trips between a certain stop pair. The sum of the whole probability flow matrix equals to 1. Heat maps are adopted to visualize those probability flow matrices. The darker the cell's color is, the higher the probability is.

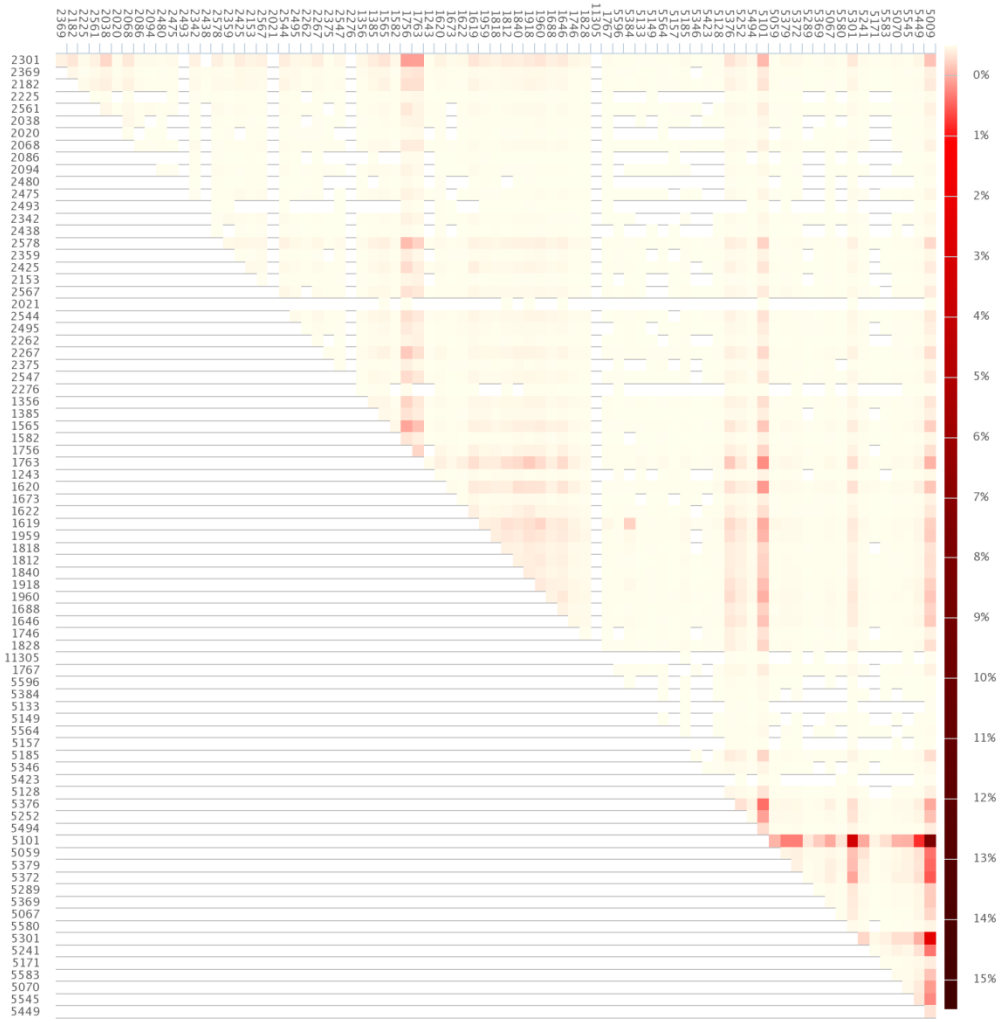


Figure 8 Probability Flow Matrix Produced by the IPF method

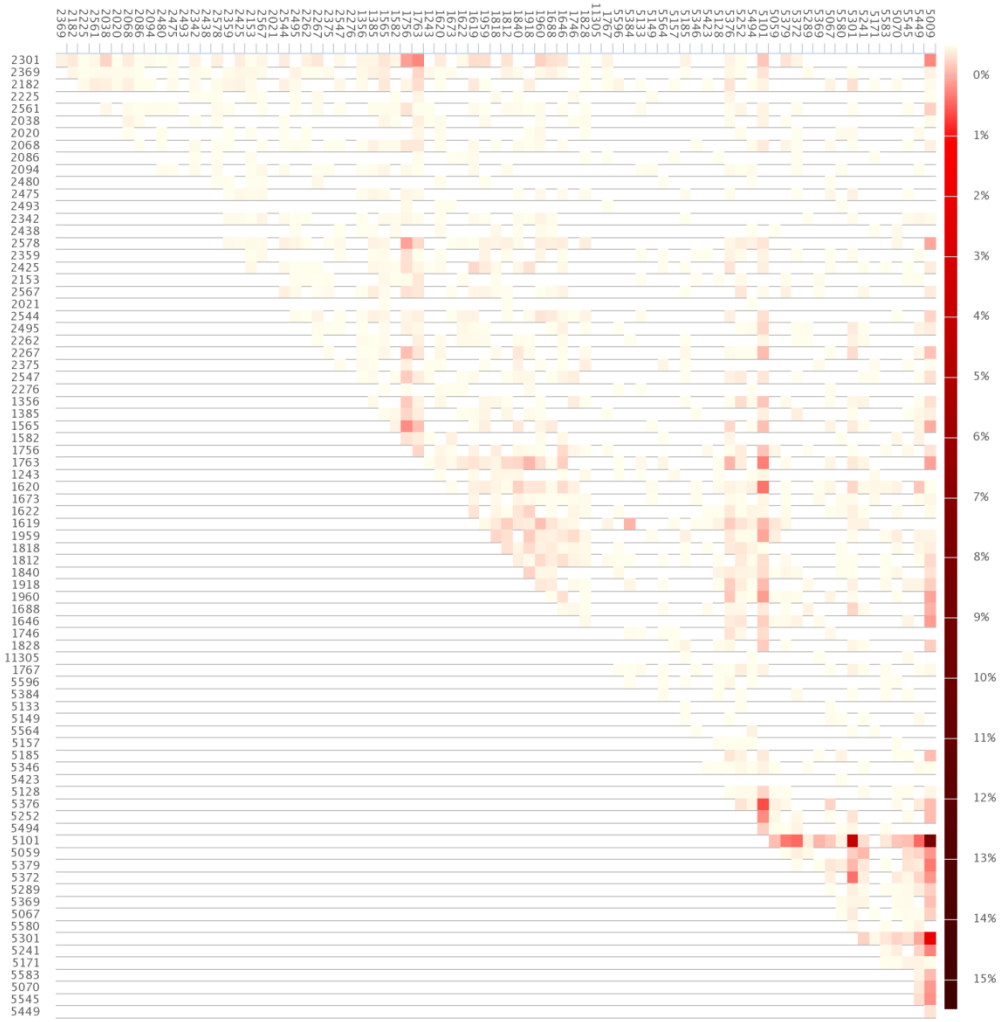


Figure 9 Probability Flow Matrix Produced by the IPF-IB method

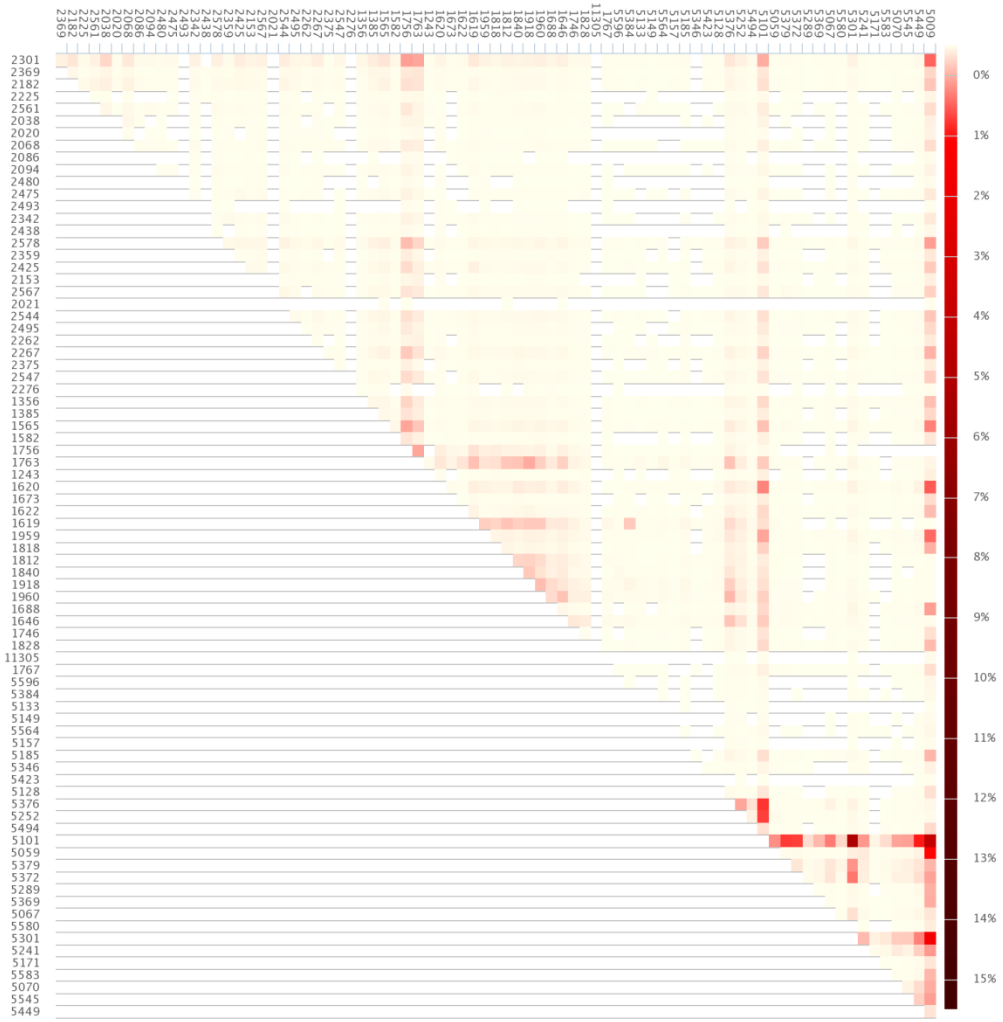


Figure 10 Probability Flow Matrix Produced by the Li and Cassidy method

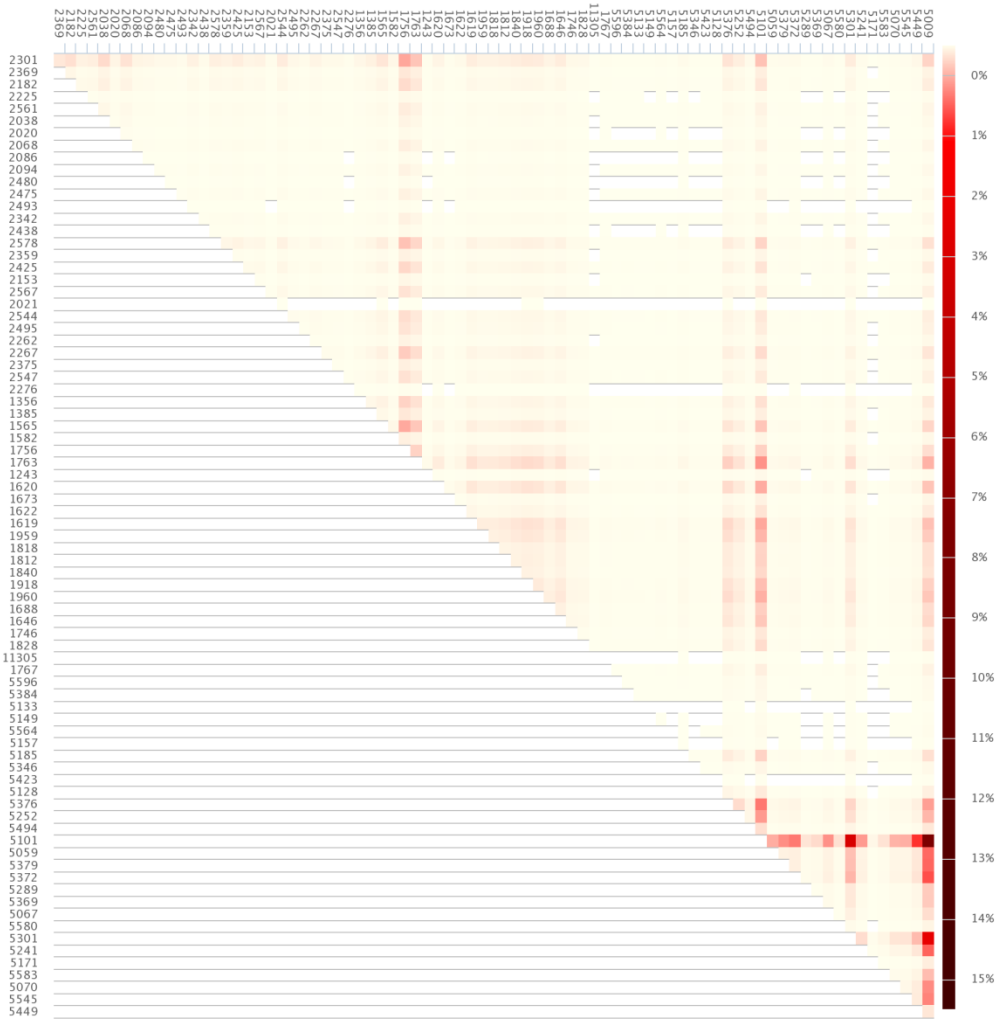


Figure 11 Probability Flow Matrix Produced by the Markov Model method

4.2.2 Result Analysis

Similar travel patterns are reflected by the heat maps of the probability flow matrices estimated by the four methods mentioned above.

Starting from the first stop 2301 of the Route 1 in the westbound direction, the stop 1756 and the stop 1763 are the first two popular stops where passengers would like to get off. These two stops are in the east of the Downtown area and

near to the LRT's Churchill Station. Passengers alight at these two stops are mainly come from the stop 2301, the stop 2369 and the stop 2182 in or near to the Capilano Transit Center in the beginning of the Route 1, the stop 2578 and its downstream stops.

The stops from 1763 to 1646 are in the Downtown area. These stops are more attractive for passengers inside this area than passengers who get on at upstream stops.

The stop 5376 and the stop 5252 which are between the Downtown area and the Jasper Transit Center are also attractive, especially for passenger from the stops in the Downtown areas.

The stop 5101 in the Jasper Transit Center is one of the busiest stops along the Route 1. The probability for passengers to make a trip which ends here is relatively high no matter which upstream stop they board at. Besides, this stop is the one where the most trips start from.

The stop 5301 in the Meadowlark Transit Center attracts passengers too, but passengers who alight at this stop mainly come from stops between the Jasper Transit Center and the Meadowlark Transit Center. The proportion of trips starting from this stop is larger than adjacent stops and most of them end at the last stop 5009.

The stop 5009 in the West Edmonton Mall Transit Center is the last stop of the Route 1. It is the most attractive stop of the Route 1 in the westbound

direction. More than one-fifth trips end at this stop and passengers mainly come from the stops in the Capilano Transit Center, the Downtown Area, the Jasper Transit Center and its downstream stops.

4.2.3 Comparison of Methods

Although they produce similar estimations, these four methods have their own strengths and weaknesses. The overall fitness measure is used to determine how these estimations fit the sample of trips in the westbound direction in the Midday time period. Table 4 lists the comparison results of the IPF method, the IPF-IB method, the Li and Cassidy method, and the Markov Model method.

Table 4 Comparison of IPF, IPF-IB, Li and Cassidy and Markov Model Method

Method	Overall Fitness	Computation Time
IPF	2.515	~ 1.35 seconds
IPF-IB	2.369	~ 3 minutes
Li and Cassidy	2.322	~ 2 hours
Markov Model	2.604	~ 0.02 seconds

In the term of this overall fitness measure, the Li and Cassidy method is better than the other three methods. It has the smallest (best) fitness measure of 2.322. However, the Li and Cassidy method is very time-consuming in this research. It cost more than 2 hours to get the result. This method is supposed to be computationally efficient because the number of iterations required for this

method is fixed theoretically. The value of α_a and α_b for major stops and minor stops, and the range of minimum riding distance are specified beforehand. However, the selection of major stops and minimum riding distances may vary from route to route. This research uses an iterative process to test different major stop groups from total 80 stops in order to determine the one which produces the best estimations, instead of using predefined major stops in the [11]. It requires a long computation time.

The Markov Model method is the fastest method among the four methods. It can produce the result within about 0.02 seconds because it develops a closed-form equation to infer passenger alighting probabilities based on the Markov Model. But its overall fitness measure value 2.604 is the biggest (worst).

The overall fitness measure values of estimations produced by the IPF method and the IPF-IB method are between that of the Li and Cassidy method and the Markov Model method. The result of the IPF-IB method is similar to the Li and Cassidy method's output, while the result of the IPF method is close to that of the Markov Model method. From their overall fitness measure values, the IPF-IB method indeed improves the IPF method. Figure 12 shows the evolvement of the overall fitness measure value in each iteration of the IPF-IB method. In this research, the IPF-IB method takes 162 iterations to reach the final results. It is not surprising that the IPF-IB method cost more computation time than the IPF method, but the IPF-IB method is much faster than the Li and Cassidy method. Moreover, the overall fitness measures of results are tended to be stable after 30 iterations according to Figure 12. It may imply that the convergence threshold of

the IPF-IB method in this research is too strict. If a looser threshold is used in the IPF-IB method to finish the calculation within 40 iterations, there is no doubt that the performance of this method will be improved significantly.

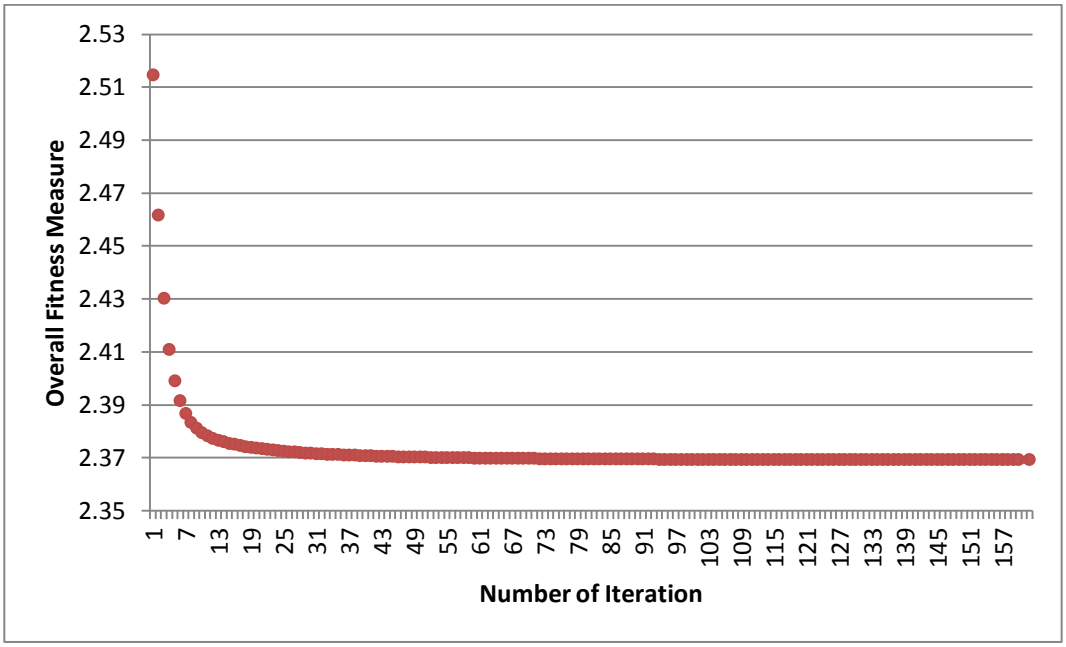


Figure 12 Overall Fitness Measures of Different Iterations in the IPF-IB method

In this case study, however, the overall fitness measures of these four methods do not show significant differences between each other. The differences between these values vary from 0.047 to 0.282, which means that estimated average loads derived from OD estimations produced by these four methods in this case study may deviate from observed average loads at a very similar level, i.e., 2 or 3 passengers. From the transit planning perspective, this deviation is tolerated.

Regarding the complexity of methods, the IPF method and the IPF-IB method is straight-forward and easy to implement, while the Li and Cassidy

method and the Markov Model method is not very intuitive because these two methods need to develop and calibrate complex models.

Considering all factors discussed above, in some cases which do not require careful estimations, the Markov Model method may be the best option because of its advantage in computation. If the accuracy of estimations matters in practical applications, then IPF-IB method might be more suitable.

4.3 Stop Grouping

Table 5 presents stop grouping in the direction from the Capilano Transit Center to the West Edmonton Mall Transit Center in the Midday period. This research investigates the result from 5-group configuration to 15-group configuration. Stop indexes in the table are the last stop index of each group. The stop grouping method is applied to estimations produced by the four methods mentioned above. Those stop indexes which do not appear in all grouping results are in red color, otherwise in black color.

Table 5 Stop Grouping Based on Estimations Produced by the Four Methods

***m*-Group**	**IPF**	**IPF-IB**	**Li and Cassidy**	**Markov Model**
5	*61, 64, 65, 79, 80*	*49, 64, 65, 79, 80*	*49, 64, 65, 79, 80*	*61, 64, 65, 79, 80*

***m*-Group**	**IPF**	**IPF-IB**	**Li and Cassidy**	**Markov Model**
6	*49, 61, 64, 65, 79, 80*	*30, 49, 64, 65, 79, 80*	*30, 49, 64, 65, 79, 80*	*49, 61, 64, 65, 79, 80*
7	*30, 49, 61, 64, 65, 79, 80*	*30, 34, 49, 64, 65, 79, 80*	*30, 49, 61, 64, 65, 79, 80*	*30, 49, 61, 64, 65, 79, 80*
8	*1, 30, 49, 61, 64, 65, 79, 80*	*30, 34, 49, 61, 64, 65, 79, 80*	*1, 30, 49, 61, 64, 65, 79, 80*	*1, 30, 49, 61, 64, 65, 79, 80*
9	1, 30, 34, 49, 61, 64, 65, 79, 80	1, 30, 34, 49, 61, 64, 65, 79, 80	1, 30, 34, 49, 61, 64, 65, 79, 80	1, 30, 34, 49, 61, 64, 65, 79, 80
10	1, 30, 32, 34, 49, 61, 64, 65, 79, 80	1, 30, 32, 34, 49, 61, 64, 65, 79, 80	1, 30, 32, 34, 49, 61, 64, 65, 79, 80	1, 30, 32, 34, 49, 61, 64, 65, 79, 80
11	1, 30, 32, 34, 38, 49, 61, 64, 65, 79, 80	1, 30, 32, 34, 38, 49, 61, 64, 65, 79, 80	1, 30, 32, 34, 38, 49, 61, 64, 65, 79, 80	1, 30, 32, 34, 38, 49, 61, 64, 65, 79, 80

m-Group	IPF	IPF-IB	Li and Cassidy	Markov Model
12	1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 79, 80	1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 79, 80	1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 79, 80	1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 79, 80
13	*1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 78, 79, 80*	*1, 15, 30, 32, 34, 38, 49, 61, 63, 64, 65, 79, 80*	*1, 30, 32, 34, 38, 47, 49, 61, 63, 64, 65, 79, 80*	*1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 78, 79, 80*
14	*1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 74, 78, 79, 80*	*1, 8, 15, 30, 32, 34, 38, 49, 61, 63, 64, 65, 79, 80*	*1, 30, 32, 34, 38, 47, 49, 61, 63, 64, 65, 78, 79, 80*	*1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 74, 78, 79, 80*
15	*1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 72, 74, 78, 79, 80*	*1, 8, 15, 30, 32, 34, 38, 47, 49, 61, 63, 64, 65, 79, 80*	*1, 30, 32, 34, 38, 47, 49, 61, 63, 64, 65, 74, 78, 79, 80*	*1, 30, 32, 34, 38, 49, 61, 63, 64, 65, 72, 74, 78, 79, 80*

The four methods used to estimate transit passenger OD matrix produce very similar group configurations. Especially in the 9-group, 10-group, 11-group, 12-group configurations, the group selections are the exactly same. When the

number of stops in a group configuration is greater than 12, grouping results of four estimations become slightly different from each other, but they are still have most stops in common. This phenomenon reflects estimations made by the four different methods are similar to each other from another angle somehow. Especially for the IPF method with the null base matrix and the Markov Model method, they generate identical group configurations. Moreover, it also suggests the group configurations which have 9 stops to 12 stops may grasp the essential travel patterns the most for the sample this research studied.

Figure 13 depicts the 9-group configuration along the Route 1 from the Capilano Transit Center to the West Edmonton Mall Transit Center. This group configuration is very reasonable from the point view of land use. The Group A is the beginning of this route in the Capilano Transit Center. The Group B covers a large residence area. Stops in this group serve the communities in this area along the route. The Group C has a different land use characteristics from its preceding groups. It is in Edmonton's Chinatown area and it has many parking lots and the first LRT station which is adjacent to Route 1. The Group D is the Downtown area where there are many commercial buildings, including shopping malls, restaurants, banks and so on. The Group E is very similar to the Group B, which covers a large area of communities. The Group F is a small commercial area next to communities in the Group E. There are some appliances stores and many finance agencies which provide cash loan, exchanges and other related services in this area. The Group G is the stop in the Jasper Transit Center. It is a very important stop on the Route 1. The Group H mainly consists of residence areas.

Some public agencies like hospitals, libraries and churches are also in this group. The Group I is the last stop of the Route 1 in the West Edmonton Mall Transit Center.

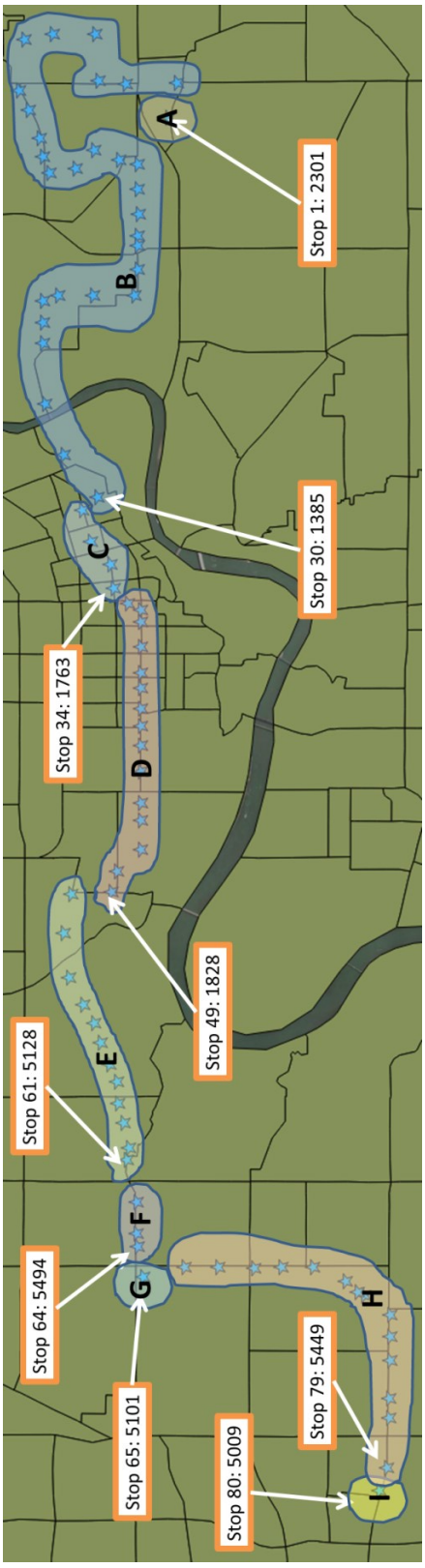


Figure 13 The 9-Group Configuration

Figure 14 is the heap map of the probability flow matrix produced by the IPF-IB method for the 9-group configuration. Table 6 lists the Groups which have large demand for the Route 1 from the perspective of probability flow. More than 50% trips are related to the Group B, D, G, H and I.

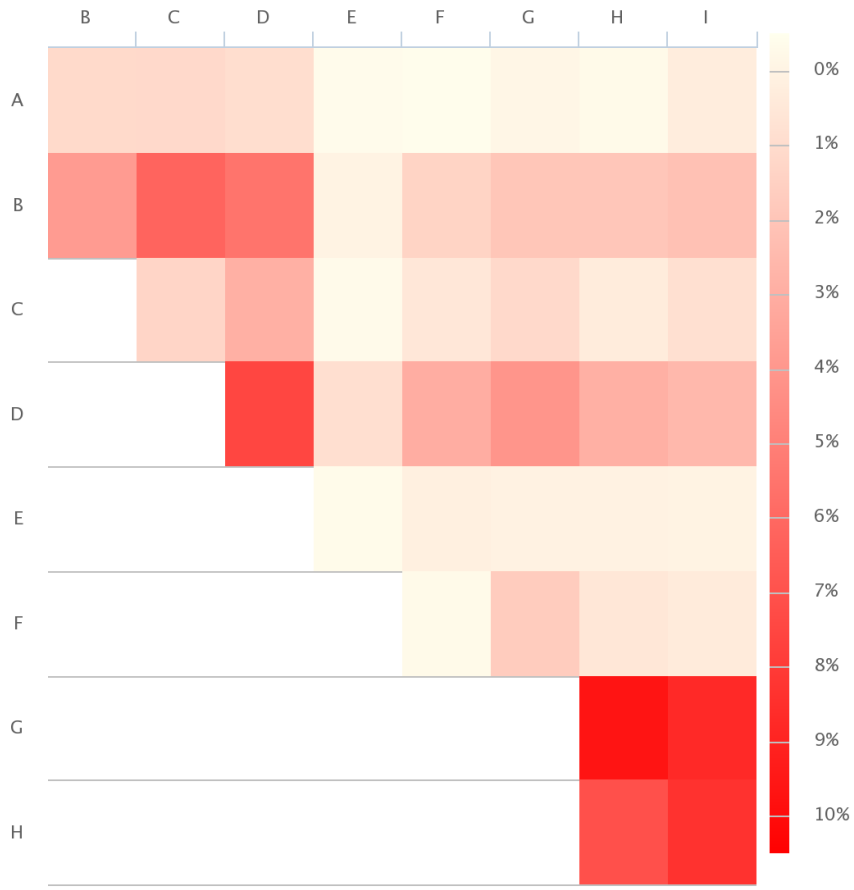


Figure 14 Probability Flow Matrix of the 9-Group Configuration

Table 6 Top Probabilities Flow between Groups

Groups	Probability Flow
G to H	9.21%
G to I	8.38%
H to I	8.03%
D to D	7.27%
H to H	6.08%
B to C	6.07%
B to D	5.47%
Total	**50.51%**

Trips beginning from the Group G occupy 17.59% of total trips. The origin of those trips is the stop in the Jasper Transit Center, where many passengers may transfer. Destinations of these trips are the Group H and the Group I. Both of these two groups seem very attractive. The Group I is the West Edmonton Mall Transit Center and the Group H between them consists of many communities. The travel time of the Route 1 from Group G to Group I is 13 minutes in off-peak hours, which is the shortest one compared with other routes between the same origin and destination. Besides, Route 1 is also a little more frequent than others starting from the Jasper Transit Center in the same direction.

These reasons make the Route 1 a good option for passengers to travel from the Group G to the Group H and the Group I, as well as from the Group H to the Group I, in which case 8.03% of total trips occur.

7.27% of total trips happen inside the Group D, the Downtown area. There are several LRT stations and many commercial buildings inside this area. These facilities produce and attract passenger flows. Therefore, the Group D itself has lots of demand for Route 1.

Trips inside the Group H, from the Group B to the Group C and from the Group B to the Group D contribute 6.08%, 6.07% and 5.47% respectively. The Group H and the Group B are two large residence areas along the Route 1 and imply large demand for travels. The Group C is an attractive destination because it is in Edmonton's Chinatown and it has the Churchill LRT Station and many parking lots. Passenger boarding or alighting at this stop group may have a good chance of transferring from or to other travel modes.

Figure 15 shows the alighting probability for each stop group. The x-axis of those charts lists names of groups and the y-axis expresses the alighting probability from upstream groups. The alighting probability is obviously affected by the travel distance and attraction of destinations. Generally speaking, passengers on the Route 1 tend to take short trips. Their alighting probabilities at destinations which are close to their origin are higher than those at destinations which are far away.

Passengers who get on at the stops in the Group A, B and C have about 20% to 30% probability to get off in the upstream groups of the Group D and have about 10% probability to get off in the downstream groups of the Group D. The probability for passengers alighting at the same group in the Group B and C is between 15% and 20%.

The Group D has its own specific alighting probability pattern. Passengers boarding in the group tend to take trips within the group. More than 30% trips happen in this area.

The Group E, F, G and H also reflect the facts that passengers prefer shorter trips than longer ones according to the alighting probabilities in these groups. Alighting probabilities of passengers from these groups vary from 27% to 54%, which are relatively higher than the alighting probabilities of upstream groups because these groups are closer to the end of this route.

Regarding the destination groups, the Group D is attractive for passengers from its upstream. The Group E might be the least attractive group along the route. The Group G attracts passengers from its upstream groups and the alighting probability at this group increases as the boarding group is closer to it. Passengers are more likely get off at the Group I than at the Group H if they get on before the Group D. After the Group D, passengers have similar alighting probability for the Group H and the Group I.

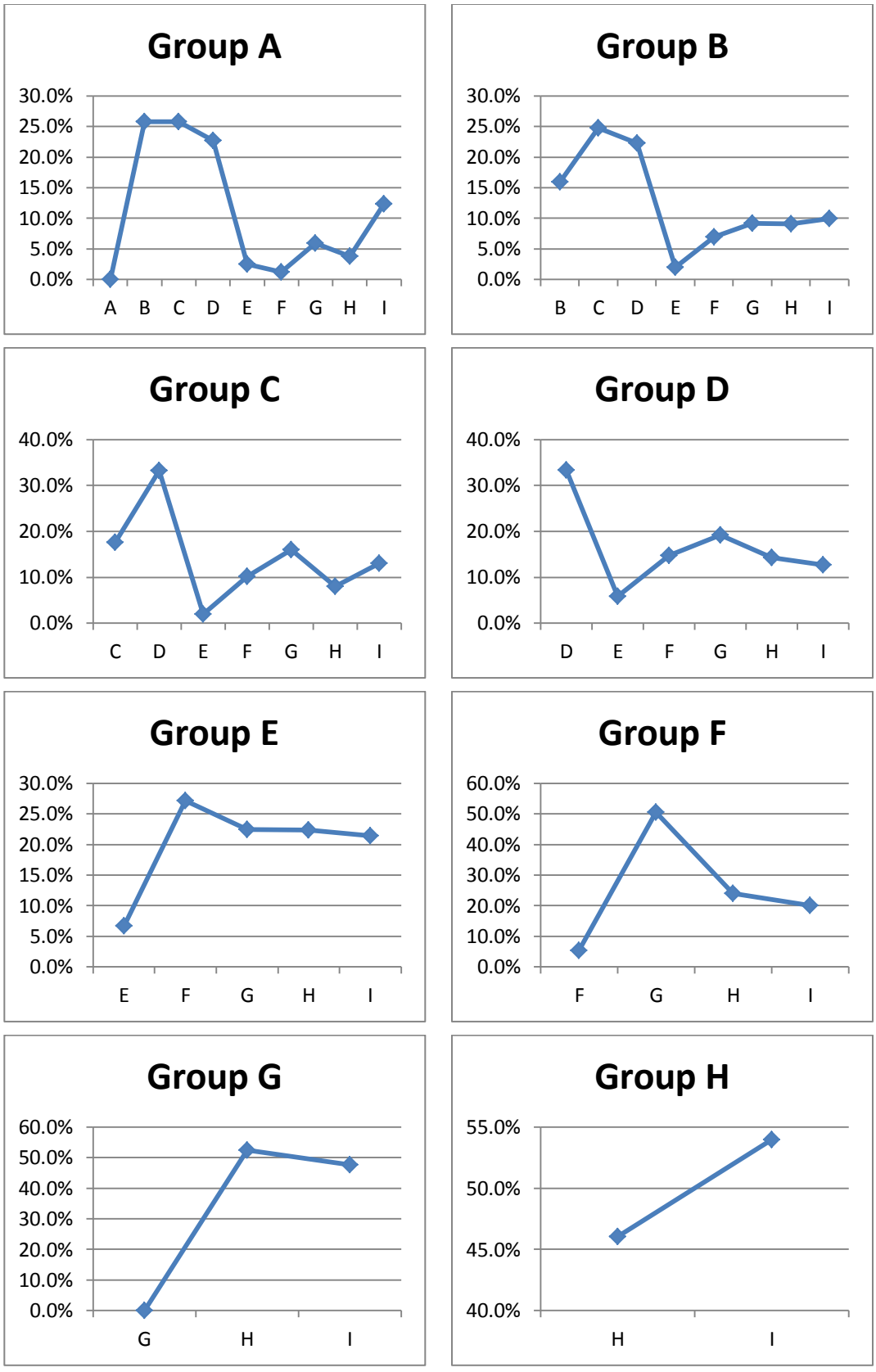


Figure 15 Alighting Probability between Groups

4.4 Summary of Results and Discussions

After applying data preprocessing, trips extracted from the stop-based APC data are grouped into 12 homogenous trip groups according to their directions and time-of-day periods. The trip group from the Capilano Transit Center to the West Edmonton Mall Transit Center in Midday has the maximum number of trips. All reviewed transit OD estimation methods are applied to this trip group.

Those four estimation methods, the IPF method, the IPF-IB method, the Li and Cassidy method, and the Markov Model method have different merits and demerits regarding the overall fitness measure and the computation time, but they can produce very similar results in the case study in this research. Therefore, the Markov Model method, which can finish the estimation very fast, is suggested to be used in most cases except those require more accurate estimations for some reasons. In those cases, the IPF-IB method can be considered.

The grouping result produced by the bus stop grouping method is reasonable from the perspective of land use, and synthesizes essential travel flow patterns along the route in this case study. The group configuration identifies the characteristics of passenger travel behaviours in the route studied in this research and areas which have high demand.

CHAPTER 5. CONCLUSIONS AND FUTURE WORKS

This chapter gives a summary of this research and discusses the limitations. This chapter also puts forward some thoughts for future works on transit passenger OD trip estimations from APC data provided by ETS.

5.1 Research Summary and Limitations

Four methods, the IPF method, the IPF-IB method, the Li and Cassidy method and the Markov Model method, to estimate a route-level transit passenger OD matrix using boarding and alighting counts are investigated and analyzed from their fitness and computation performance in this thesis.

The IPF method is a traditional method to estimate OD matrices. Its idea is to match sums of rows and columns of a matrix to observed boarding and alighting counts respectively. It requires a base OD matrix. In this research a null base OD matrix is used since there is no existing base OD matrix. Although a null base OD matrix is acceptable [7], it introduces some biases on the result because its assumption of equal flow for each trip is not very realistic. The IPF-IB method uses an iteratively improved base matrix to overcome the bias caused by the null base matrix. This iterative process improves the fitness of the estimated result significantly.

The Li and Cassidy method doesn't require a base OD matrix. It develops a model to estimate the number for passengers who get on at major stops to alight

at a certain stop. In this research, an iterative method is proposed to test different major stop groups in order to get the best-fit OD estimation. The Markov Model method doesn't require a base OD matrix either. It employs the Markov chain model to express the alighting probability for a passenger to alight at a certain stop and develop a closed-form equation which takes Markov transition probabilities as its parameters to calculate the alighting probability for each stop pair.

These four methods are applied to the APC data of the Route 1 collected by ETS from 30 June, 2014 to 29 August, 2014. The Li and Cassidy method can generate the estimation which has the best fitness but it consumes the longest computation time, more than 2 hours, because this method tries to find the best combination of the major stops groups, the minimum riding distance and the auxiliary parameters α for major stops and minor stops. The Markov method has the best computation performance, about 0.02 seconds, while the fitness of this estimation is not as good as others. The IPF method can get the result in this research within few seconds and its fitness of the result is similar to the Markov Model method's result. And the estimation of the IPF-IB method is similar to the one produced by the Li and Cassidy method, but it can finish its computation within few minutes, which is acceptable. The overall fitness measure analysis indicates it is possible to improve the computation performance of the IPF-IB method further. Although those methods have pros and cons, their results do not have significant differences. The Markov Model method is recommended for most cases. In case some practical application cares the accuracy of estimations

very much, the IPF-IB method might be a good option with considering the overall fitness measure and the computation time together.

In addition, the method of grouping stops is reviewed and applied to the APC data of the Route 1 in order to reduce the size of the OD matrix and figure out the critical travel pattern along this route. An identical 9-group configuration is identified based on the results produced by different estimation methods. This stop group is reasonable from the perspective of land use and reflects important flow patterns. Based on this group configuration, the stops in Jasper Transit Center and its downstream, and the stops in the Downtown area have outstanding demand for Route 1 in the direction from the Capilano Transit Center to the West Edmonton Mall Transit Center.

However, there are some limitations in this research. The analysis in this thesis is based on estimations from one single route's APC data in one time-of-day period. It would be beneficial to repeat this analysis on data of different routes and different time periods. Besides, there is no true OD matrix of the Route 1 that can be used as a benchmark to evaluate the estimations generated by different methods.

5.2 Future Works

Every estimation method has its own assumptions and inputs. It is worthy to conduct a comprehensive analysis on the relationships between them and estimations results. Some potential future works are listed as follows:

- Figure out the relationship between the sample size and the result's fitness.

- Develop an efficient numerical method to determine major stops in the Li and Cassidy method.
- Refine the Markov model in the Markov Model method to improve its estimation's fitness.
- Investigate the sensitivity of different dissimilarity measure (e.g., chi-square testing) in the stop grouping method.
- Check the reasonability of transit operation periods defined by ETS.

Moreover, it is valuable to extend this research from a single route to the whole transit network by considering passenger transfer [31] so that it is possible to depict the demand and travel patterns for the whole transit network in Edmonton from APC data.

Finally, once a method or a combination of several methods are proven suitable for the entire ETS network, it might be helpful to develop a product to analyze the whole APC data source and provide suggestions so that the transit agency may conduct planning, manage services and monitor operations in a better way.

REFERENCES

[1] H. J. Van Zuylen and L. G. Willumsen, "The most likely trip matrix estimated from traffic counts," *Transportation Research Part B: Methodological,* vol. 14, no. 3, pp. 281-293, 1980.

[2] Y. Gur, "Estimating trip tables from traffic counts: Comparative evaluation of available techniques," *Transportation Research Record,* no. 944, pp. 113 -117, 1983.

[3] M. E. Ben-Akiva, P. P. Macke and P. S. Hsu, "Alternative methods to estimate route-level trip tables and expand on-board surveys," *Transportation Research Record,* no. 1037, pp. 1-11, 1985.

[4] D. Lu, "Route Level Bus Transit Passenger Origin-Destination Flow estimation Using APC Data: Numerical and Empirical Investigation," in *MS thesis*, Columbus, Ohio, USA, The Ohio State University, 2008.

[5] P. Delle Site and F. Filippi, "Service optimization for bus corridors with short-turn strategies and variable vehicle size," *Transportation Research Part A: Policy and Practice,* vol. 32, no. 1, pp. 19-38, 1998.

[6] A. Tirachini, C. E. Cortés and S. R. Jara-Díaz, "Optimal design and benefits of a short turning strategy for a bus corridor," *Transportation,* vol. 38, no. 1, pp. 169-189, 2011.

[7] M. R. McCord, R. G. Mishalani, P. Goel and B. Strohl, "Iterative proportional fitting procedure to determine bus route passenger origin-destination flows," *Transportation Research Record: Journal of the*

Transportation Research Board, vol. 2145, no. 1, pp. 59-65, 2010.

[8] M.-P. Pelletier, M. Trépanier and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies,* vol. 19, pp. 557-568, 2011.

[9] M. A. Munizaga and C. Palma, "stimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies,* vol. 24, pp. 9-18, 2012.

[10] A. Cui, "Bus passenger Origin-Destination matrix estimation using automated data," in *M.S. Thesis*, Cambridge, MA, Massachusetts Institute of Technology, 2006.

[11] Y. Li and M. J. Cassidy, "A generalized and efficient algorithm for estimating transit route ODs from passenger counts," *Transportation Research Part B: Methodological,* vol. 41, no. 1, pp. 114-125, 2007.

[12] P. G. Furth and D. S. Navick, "Bus route OD matrix generation: Relationship between biproportional and recursive methods," *Transportation Research Record,* vol. 1338, pp. 14-21, 1992.

[13] "Future Transit - Bus and LRT," Edmonton Transit System, [Online]. Available: http://www.edmonton.ca/transportation/ets/future-transit.aspx. [Accessed 13 01 2015].

[14] P. Furth, B. Hemily, T. H. J. Muller and J. Strathman, "Using archived AVL-APC data to improve transit performance and management,"

Transportation Research Board, Washington, D.C, 2006.

[15] P. G. Furth, J. G. Strathman and B. Hemily, "Part 4: Marketing and Fare Policy: Making Automatic Passenger Counts Mainstream: Accuracy, Balancing Algorithms, and Data Structures," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 1927, no. 1, pp. 205-216, 2005.

[16] Y. Ji, R. G. Mishalani and M. R. McCord, "Estimating Transit Route OD Flow Matrices from APC Data on Multiple Bus Trips Using the IPF Method with an Iteratively Improved Base: Method and Empirical Evaluation," *Journal of Transportation Engineering,* vol. 140, no. 5, 2014.

[17] R. G. Mishalani, Y. Ji and M. R. McCord, "Empirical evaluation of the effect of onboard survey sample size on transit bus route passenger OD flow matrix estimation using APC data," *Transportation Research Record,* vol. 5, no. 2246, pp. 64-73, 2010.

[18] B. Li, "Markov models for Bayesian analysis about transit route origin–destination matrices," *Transportation Research Part B: Methodological ,* vol. 43, no. 4, pp. 301-310, 2009.

[19] L. M. Bregman, "he relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics,* vol. 7, no. 3, pp. 200-217, 1967.

[20] B. Lamond and N. F. Stewart, "Bregman's balancing method,"

Transportation Research Part B: Methodological, vol. 15, no. 4, pp. 239-248, 1981.

[21] J. Simon and P. G. Furth, "Generating a bus route o-d matrix from on-off data," *Journal of Transportation Engineering,* no. 6, pp. 583-593, 1985.

[22] S. E. Fienberg, "An iterative procedure for estimation in contingency tables," *The Annals of Mathematical Statistics,* vol. 41, pp. 907-917, 1970.

[23] S. Tsygalnitsky, *Simplified methods for transportation planning,* Cambridge: Massachusetts Institute of Technology, 1977.

[24] M. R. McCord, R. G. Mishalani and X. Hu, "Grouping of Bus Stops for Aggregation of Route-Level Passenger Origin-Destination Flow Matrices.," *Transportation Research Record: Journal of the Transportation Research Board,* vol. 2277, no. 1, pp. 38-48, 2012.

[25] L. Le Cam and G. L. Yang, Asymptotics in statistics: some basic concepts, Berlin: Springer, 2000.

[26] J. Zhao, *The planning and analysis implications of automated data collection systems: rail transit OD matrix inference and path choice modeling examples,* Cambridge, MA: Massachusetts Institute of Technology, 2004.

[27] Edmonton Transit System, "The way we move using public transit: 2011-2013 ETS Business Plan," Edmonton, 2011.

[28] "ETS Map - Day Service," Edmonton Transit System, 30 11 2014. [Online]. Available:

http://www.edmonton.ca/transportation/transit/ETS_Day_Map_November_2014.pdf. [Accessed 9 12 2014].

[29] "Map of Route 1," Edmonton Transit System, 31 8 2014. [Online]. Available: http://webdocs.edmonton.ca/transit/route_schedules_and_maps/future/RT001.pdf. [Accessed 9 12 2014].

[30] F. Pukelsheim and B. Simeone., "On the iterative proportional fitting procedure: Structure of accumulation points and L1-error analysis," 2009.

[31] "Population Forecasts," City of Edmonton, 1 4 2014. [Online]. Available: http://www.edmonton.ca/business_economy/demographics_profiles/population-forecasts.aspx. [Accessed 10 12 2014].

[32] Q. Zhang, "OD Flow Estimation for a Two-Route Bus Transit Network Using APC Data: Empirical Application and Investigation," The Ohio State University, 2008.