Origin-Destination Trip and LRT Ridership Estimation with New Data Source

By

Difei He

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

Transportation planning are important for improving traveler's efficiency and saving energies to provide a more sustainable community. Traditional methods of conducting the transportation planning origin-destination estimation and multi-modal analysis are heavily relying on the labour-intensive data collection that are no longer suitable for today's increasing demand of travel needs. That being said, the rapid development of new technologies in telecommunication networks is producing large amounts of network data regarding how people and their devices move around in the city. In contrast to traditional GPS data that required additional geographical sensors and applications to record the information, network data has the advantage of high market penetration rates, low costs, and daily collected geographical information when considering urban travel behaviour analysis. The geographical information embedded in the network data offers researchers the potential to investigate travel mobility behaviour. However, due to the noise and spatial/temporal sparsity of network data, extracting mobility information, such as transport mode, from these data is challenging. This thesis proposes a complete architecture of transport mode detection based on the network data to monitor the Light Rail Transit ridership during daily use and to estimate the ridership and origin-destination matrices from an "easy-to-detect" transport mode, like the LRT. A hybrid heuristic method that combines a time-window based method and a pattern-based method is proposed to process the raw network data, followed by a binary logit model to estimate the probability of one candidate trip being an LRT trip. The statistical results from the network data analysis are validated by third party data reported by the City of Edmonton that shows passengers boarding and alighting at each station. Although the performance of the proposed methods lacks prior analysis, owing to the absence of ground truth, the results in this study are analyzed based on the prior knowledge and intuitive understanding of the City’s LRT system operation. Finally, this study reviews the current research gaps within the transportation field regarding data cleaning methods, mode detection models, and bias issues.

ACKNOWLEDGEMENT

I am deeply indebted to my supervisor, Dr. Tony Qiu, for his constant encouragement and guidance, without which I could not have completed my studies. I would also like to thank Dr. Tae. J Kwon, Dr. Douglas Tomlinson, and Dr. Emily Grise, for serving on my committee. Chenhao Wang and Jiangchen Li have made helpful suggestions during the course of this study, for which I am grateful. I would especially like to thank Lian Gu, who supported me through this process. Thanks also to my colleagues in Dr. Qiu's research group for the helpful discussions and collaboration.

I gratefully acknowledge the support from the City of Edmonton, who provided the data to support my research.

Finally, I would like to express my unending appreciation to my parents, who have supported me throughout my life journey and provided me with the opportunities to broaden my horizons.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

GTFS	General Transit Feed Specification
APC	Automatic Passenger Count
AFC	Automated Fare Collection
CDR	Call Detailed Record
BTS	Base Transceiver Station
BSC	Base Station Controller
LA	Location Area
TAZ	Traffic Analysis Zone
RBH	Rule-Based Heuristic
CoO	Cell of Origin
LRT	Light Rail Transit

CHAPTER 1. INTRODUCTION

This chapter presents the background of the study using network data to conduct multi-modal transportation planning analysis. The limitations of traditional on-board survey method is also described and the research motivation, research objectives, and the structure of the thesis are also presented.

1.1 Background

Transportation planning is an important field that can impact on the efficiency and sustainability of commuters' daily experience and routine. Government decision-makers are keen to understand how people travel through the city to facilitate infrastructure investment decisions. The four-step model, as one of the most prevalent models in the transportation field, has been utilized to help transportation engineers and researchers quantify the travel behaviour in a given study region. It is a systematic view of how residents regularly travel around a city, providing a strategic view of how city builders can improve the system. The model includes trip generation, trip distribution, mode split, and traffic assignment. Through these four steps, urban movements are categorized into several types of trips that represent the overall city travel patterns.

However, human travel behaviour is highly complex, and collecting and classifying the travel data can be extremely difficult and tedious. Over the past several decades, researchers have implemented various methods of collecting travel data that can be used to supplement this transportation planning model(*1*). GPS-enabled devices, travel diary surveys, and traditional video surveillance data, are all methods that can confer travel patterns to a degree. From the 1990s(*1*), growing cell phone use has given researchers a new way of tracking passengers daily movement. The data provided by cell phone use have facilitated better understanding of the way people travel around cities, helping city managers seeking to improve the transportation system to meet increasing travel demands.

Among all the data collected by various technologies, the travel trajectories, defined as a set of location records with the corresponding timestamp to infer the human movements over a certain time period, have been utilized by researchers and engineers to monitor urban dynamics. Before the big data era, transportation experts used telephone surveys, mail-in travel diaries, and point-based observatory equipment to estimate daily traffic (*2*). The traditional four step-model heavily relies on trip diaries to reflect travel patterns and continues to be one of the most common types of travel trajectories. However, these traditional methods are time-consuming, labour-intensive, and costly, with results only updated once every five or ten years. Today, community development and travel demands are growing rapidly and a newer, faster way of evaluating travel patterns, and thus needs, is required.

More recently, people have tried a variety of new technologies to monitor human movements across the city. Point-based technologies like video cameras and loop detectors are stationary data collectors are one method but only cover small areas, limited by their fixed locations. As such, the collected data cannot accurately represent the movements of the whole transportation network. Probe-vehicle type technologies like GPS data and network data, on the other hand, are location-information devices that can be carried with individual passengers or vehicles across an entire journey. A key drawback here is that these technologies are restricted by their market penetration rate and, as such, the sampling size is unlikely to be representative of the whole population.

Network data, as one of the many new data sources, can play an important role in supplementing the traditional traffic analysis tools. Network data is collected by mobile phone operators for non-transportation purposes like billing and network operation, etc. With no accurate geographical location information given, the data framework informs the connected cell tower, current timestamp, and billing activities. Due to the technological limitations of the cell network framework, network data are commonly utilized in macro-level travels including Origin-Destination (OD) table estimation or inter-city traffic. However, long update time intervals and inaccurate location information have historically made the data difficult to be implemented for micro-level traffic movements, like traffic volumes on a designated road segment, or for transportation mode detection.

This, however, is changing and offering greater opportunity for data gathering. In the era of 4G technologies, the number of cell towers has significantly increased due to the shorter coverage distance of high-frequent signals. As a result, update intervals between two consecutive cell records have reduced from average 86 minutes down to 5 minutes(*3*). With the improvement of both temporal and spatial resolution, the network data make smaller-scale and more detailed analysis possible. The shorter update intervals can better reconstruct trip trajectories, and the higher density of cell towers provide more accurate location information.

Of course, traditional Call Detailed Record (CDR) data and sighting data can only collect trace information when subscribers are actively using the mobile phone. CDR data, normally referred as the billing records collected by the network service providers, collects timestamps and spatial reference information when subscribers are using the cellphone. Sighting data, on the other hand, uses the triangulation method from the surrounding network stations to approximately estimate the locations of subscribers. The signal strength of three closest network stations are calculated to provide the approximate latitude and longitude of the current location. We can see from previous research (*3*)(*4*)(*5*) that the two data sources are still sparse in temporal resolution, and trips that happen between two recorded data points are missed due to low update frequency. On the contrary, new 'ping' records have been introduced in the network system that regularly collect the connection information, even when mobile phones are not in use. These 'ping' records can largely make up the gap between two activities and reconstruct the real movement trajectories that CDR data and sighting data cannot capture. For the case presented here, the average update interval between two consecutive records is 5 minutes, which is considerably smaller than CDR data or sighting data. In order to support high web access speed, especially in urban areas, cell towers intricately overlap. In our case, as for most urban areas, subscribers could be covered by more than ten cell towers simultaneously.

1.2 Problem Statement

The traditional methods of information gathering for transportation planning, like household travel surveys and four-step modelling, have difficulty monitoring the increasing demands in transportation. The conventional household travel survey does not provide enough information on public transportation service for city residents, nor does it provide insight into areas for service improvement. In addition, public transport analysis is based on the household data suevey information that is collected annually and is not always representative due to its small sample sizes and relatively short time span. Therefore, a new and better way to estimate and map public transportation usage on a day-by-day basis is important for planners and decision makers.

Other methods pose difficulties in providing complete and accurate data for planners. LRT passenger boarding and alighting information is conducted by on-board surveys and the Automated Passenger Count (APC) system(*6*), both of which are time-consuming and costly. In addition, these methods are difficult to implement in large-scale sample sizes. Alternatively, researchers have explored various new data sources that are more affordable and accessible to estimate the transit ridership(*7*). Hand-held GPS equipment is one popular choice used to record travel information and monitor individual transit passengers. Its drawbacks are the expense of the equipment and privacy issues, making individual GPS trajectories less than ideal. On the other hand, network data as a passively collected data source possesses several strengths. Firstly, the data is collected from existing cell stations and servers, so additional infrastructure investment is unnecessary. Secondly, the penetration rate of network data normally exceeds 30%, making it more representative than any other data source. Lastly, network data covers full 24-hour periods so that full-day trajectories can be completely represented (*5*).

Current research, however, does not provide any systematic data processing architecture, given the spatial and temporal characteristics of the network data. A new proposed method that takes advantage of new generation network data would greatly benefit urban travel dynamics analysis since the increased temporal and spatial resolution opens up the possibility to extrapolate more traffic information. This thesis proposes a comprehensive data processing architecture for the given data that addresses this gap in the current research and, importantly, addresses the limitations and drawbacks that the data possesses.

1.3 Research Motivation

Network data have proved to be a great data source that covers people's movement over a full day(*8*). It can trace travel trajectories quickly and accurately without additional infrastructure investment and provide efficient ways to conduct transportation planning to meet the increased needs of urban travellers. This novel technology has provided new data sources to monitor the daily travel activities that could potentially be utilized in transportation modelling. As the largest mobility data source, it can potentially help to supplement traditional transportation planning methods and provide a macro-scale picture of daily passenger flows. However, it also requires a comprehensive understanding, appropriate assumptions, and suitable analytical skills to interpret. Recent research has seldom tried to utilize the network data to estimate the urbanized Light Rail Transit (LRT) passenger origin-destination matrices and passenger ridership(*9*)(*2*). This research presented in this thesis aims to find a practical way to interpret detailed travel information for travellers in an urban context.

1.4 Research Objectives

In this thesis, a novel methodology is proposed to identify Light Rail Transit passengers from the raw network data and estimate OD matrices for LRT stations. As such, the objectives are:

1. To provide greater market penetration rate samples as a supplement the household travel survey when analyzing the growing number of trips undertaken in urban areas;

2. To suggest a new way of monitoring public transportation over a longer time period that can map usage changes on different days of the week, or even in different seasons, while the method itself remains cost-effective;

3. To propose a novel network data processing algorithm that can take advantage of this new data source to provide detailed travel information that supplements traditional travel diary survey data.

1.5 Structure of the Thesis

This thesis includes five chapters:

Chapter 1 has introduced the background to multiple probe vehicle technologies and the estimation of transit ridership from big data technologies. The problem statement and research objectives have also been outlined in this chapter.

Chapter 2 is the literature review of related research. The review focuses mainly on research around network data and its application to the identification of transportation modes. The pre-processing methods and methods to identify the transportation modes are also included.

Chapter 3 outlines the methodology implemented in this study, and the study area of the LRT Capital Line within the City of Edmonton.

Chapter 4 analyses and presents the results generated from transit ridership identification.

Chapter 5 concludes the study and offers suggestions for future studies about transit ridership utilizing novel probe technologies.

CHAPTER 2 LITERATURE REVIEW

This chapter summarizes the current researches and literatures about the multi-modal transportation analysis to fill the gap of traditional household travel survey. The data processing techniques and transport mode identification methods utilizing the network data are also discussed in this section.

2.1 Background

In the era of big data and emerging technologies, researchers have access to various types of data when solving transportation problems. People use data mining tools and mathematical models to extract hidden information from the raw data. The following section describes the state-of-the-art methodology and applications of the big data technologies, including a variety of data sources, followed by a range of research work related to transport mode detection and public transit analysis. The limitations of previous research, as well as the comparison between the method presented in this thesis and other methods, are highlighted.

Researchers are continually searching for new approaches to transportation planning for future city growth and development. In the past several decades, collection of the data input for traditional transportation planning models, like the four-step model, has been difficult. Traditional methods of collecting traffic data have included travel diaries, loop detector data, manual traffic count boxes, and traffic surveillance cameras. All of these methods are labour intensive and expensive, which makes immediate updating of the four-step model difficult. However, recent probe vehicle technologies have become more prevalent, making the construction of daily travel trajectories easier.

Probe types of data like network data and GPS data have contributed significant amounts of travel trajectory information that facilitate the understanding of travel behaviour. Because all data are collected by different methods and technologies (*3*)(*5*)(*4*)(*10*)(*11*), their position accuracy, update interval, and penetration rates are different. Network data, for instance, possesses high penetration rates but low spatial and temporal accuracy. It is used mostly to generate Origin-Destination matrices, to identify the hotspots in urban areas, or to distinguish transport modes during long-distance travel. GPS data, on the other hand, has much higher geographical location accuracy and smaller update intervals. However, it normally requires an additional device to collect(*12*). In comparison, network data has a much higher market penetration rate, which can better represent macroscopic traffic dynamics.

This section aims to provide a comprehensive overview of the current transportation mode identification methods and approaches based on the various data sources. As demonstrated in other studies(*9*), the process for transport mode identification normally includes data cleaning, data segmentation, and mode inference, common in many urban mobility applications. Each step requires detailed analysis and understanding before proceeding to the next level, and this study aims to answer the following questions: 1) What are the characteristics of the data, and how does the data need to be implemented to identify the transport mode? 2) What are the existing pre-processing methods? How can researchers define and validate samples? 3) What is the current mode detection model that is being implemented? How does our model improve the reliability and accuracy in comparison with other proposed models?

GPS data provide accurate geographical locations with timestamps and are considered prevalent information sources for identifying and tracking movement trajectories when trying to overcome the limitations of traditional travel surveys(*13*). The dedicated GPS loggers and more recent GPS-enabled smartphones can be used to record travel movements for a whole day. Researchers have processed the data information to extract the important trip characteristics for transportation analysis(*14*). One common form of study using GPS data is to infer the transport mode using various methods and approaches, e.g. rule-based methods, logit models, machine learning, etc.(*9*). However, these approaches to collecting the traces of movement has required GPS loggers or GPS recording applications to be active for the entire time, which limits the scale of travel data collection.

2.1.1 GTFS (General Transit Feed Specification) Data for Identifying the Transit Ridership

From earlier literature, the GTFS data are predominantly used to estimate the transit vehicle trajectories(*15*). The GTFS data defines a common format for public transit in terms of scheduling and geographical information and is a significant source to monitor transit bus operations. The geographical information collected periodically can be used to monitor vehicle speed, bus stop on-time rate, etc. The data itself includes GPS information collected from on-board sensors to reflect the travel trajectories along the transit route. However, the data only reflects the vehicle trajectories in service while the passenger trajectories remain unknown. Researchers have tried using Automatic Passenger Count (APC) data(*16*), Smart Card data(*12*) for recording the boarding and alighting information, and other data sources to capture the passenger's behaviour(*8*).

In order to estimate the transit system operation and usage rate, Luo (*17*) has implemented a clustering method for estimating the origin-destination matrix for the transit system. This research combined the GTFS data as vehicle trajectories and the Automated Fare Collection (AFC) system as the passenger trajectories to infer the transit ridership for the Haaglanden area in the Netherlands. Gundlegård (*18*)(*19*), on the other hand, proposed a comprehensive method of estimating the transit ridership through GTFS records and APC data(*20*). Different types of data sources do bring biases to the analysis. For instance, while GTFS data are limited by the frequency and bus schedule, network data lean towards heavy cell phone users. Zhang (*21*), in his paper, has compared and contrasted the differences in monitoring the urban travel behaviour between various data sources and proposed a fusion data architecture to compliment the differences in each type of data.

In many cities where the transit system is lacking passenger recording information, transit ridership can be difficult to estimate independently from the GTFS data. Other types of data sources are often utilized to supplement the missing passenger information. On-board surveys, APC systems(*16*), and smart IC card information(*22*)(*23*) are the common methods to collect the detailed passenger information that, by themselves, all have drawbacks. On-board surveys normally can only be implemented in a small range and during limited periods. The APC system collects all the boarding and alighting information of passengers at each stop or station, but the more detailed origin-destination information is not available and further processes are required. Smart IC cards provide the most comprehensive information in terms of tracking passengers in and out of the station. However, the whole system requires large infrastructure investment and maintenance and is not available for all public transit.

2.1.2 Network data for Identifying the Transit Ridership

Only a few studies have utilized network data to detect transportation modes (*9*)(*2*). Rather, methods like map-matching algorithms and supervised learning algorithms have been the main approaches to infer the geolocation of transport networks, using mainly GPS data(*13*)(*24*). In the transportation planning four-step model, these methods can be utilized to estimate the trip generation and trip attraction information, but the mode split, and trip assignment may require additional data to support the analysis. Compared to GPS data, raw network data without any pre-processing are coarse, noisy, and sparse, making the reconstruction of travel trajectories more difficult.

Call Detailed Record (CDR) data and sighting data have been two major data sources for previous research that considered network data(*5*)(*25*). CDR data is collected by cellphone carriers when subscribers are actively using the cellphone for activities such as calling, text messaging, or internet access. Sighting data, on the other hand, is generated using the triangulation method to estimate the current location of the cell phone user. When derived from older generation of mobile technology, both CDR data and sighting data have some common issues: spatial resolution and temporal resolution(*5*).

2.2 Nature of Network data

During the last two decades, researchers have conducted various investigations to extrapolate embedded information in network data and apply the data to different transportation applications(*26*)(*20*)(*27*)(*28*). Network data, as a data source that is not primarily intended for transportation purposes, requires thorough analysis and appropriate assumptions to use it in a valid and appropriate way (*29*). Two types of network data are currently used: CDR data and sightings data. The former contains user ID, timestamps, and the cell tower location information that is channeling all activities. The latter includes the user ID, timestamps, and the estimated geographical locations resulting from the triangulation of multiple cell towers connected with any one mobile phone at the same time(*30*)(*31*). The temporal and spatial resolutions of the network data are two major aspects that impact the quality of results, and both should be evaluated in a detailed manner.

For typical network data, the smallest spatial entity is called a 'cell' that represents one or more antennae covering a defined area. A set of antennae covering one defined area forms a Base Transceiver Station (BTS) that is responsible for the communication between the cell network and a cellphone. A set of BTSs is controlled by a Base Station Controller (BSC) that manages the radio communications between them to ensure the network service is optimized (*9*)(*2*). Together, one or more BSCs and the 'cells' covered by BTSs formed a Location Area (LA) that covers a given region, as demonstrated in Figure 2.1. The base station antenna coverage and orientation are shown below, with each cell corresponding to one specific direction of the base station. Overlaps could exist between different sectors if the user was close enough to the base station. However, in this case, it is assumed that for each base station, the geolocation coverage is subdivided by three directions, where each set of antennae is responsible for 120 degrees coverage.

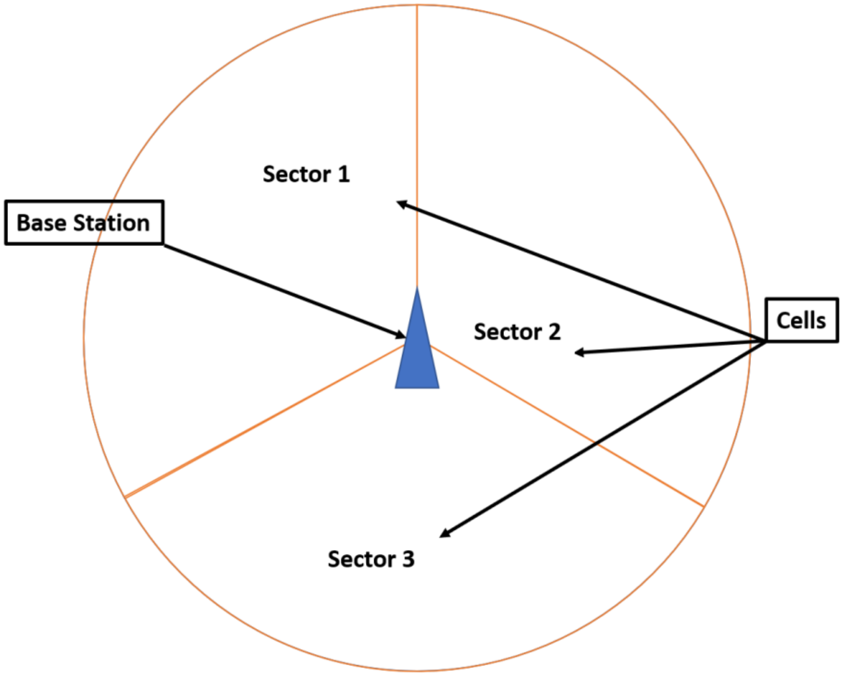


FIGURE 2.1 Base station antenna coverage and orientation example

In this context, a record or sighting of both data types is regarded as a trace of the user. In current research (*32*), the 3G network or GSM network relies on phone usage, which can be temporally sparse throughout an entire day. The trace will only be collected by the network when users are actively making use of the phone, i.e. a phone call, text message, or downloading data. Both CDR and sighting data type do not cover travel movements over the course of the entire day due to the low temporal resolution, and both are heavily dependent on the frequency of usage. The evident problem with this recording mechanism is that only the phone call, text message, or data usage activity is collected and could happen not only when the user is travelling but also when they are stationary.

Nonetheless, network subscribers, according to the International Telecommunication Union 2017 statistics, have risen to 7.7 billion internationally, with a penetration rate of 127.3%

in the developed world and 98.7% in the developing world(*3*). The potential of network data use to conduct travel behaviour analysis with its low investment costs and infrastructure support is attracting transportation researchers(*30*).

Passively generated network data, including CDR and sighting data, are characterized as temporally and spatially uncertain. For temporal resolution, network data possess a low update frequency compared with GPS data, Wi-Fi signals, and Bluetooth. The billing records are collected only when the cell phone users are calling, texting, or web browsing. Due to the data collecting mechanism, the temporal resolution can vary from person to person; the arithmetic average of the update interval means is 84 minutes, and the average inter-event time is 260 minutes, as established by previous studies(*3*). For the spatial resolution, the billing records only provide the geographical location of cell towers that are connected with the cellphones at any given time, while the sighting data estimates the approximate location of users using the triangulation of several cell towers and their signal strength to the cell phone(*33*)(*34*). It is key to note here that both types of data have significant error in estimating the users' location. It is estimated that the users' location error for billing records with cell tower information is around 1 km, and the estimated location error is 200 metres(*35*).

Therefore, since both the temporal and spatial resolution of network data have been acknowledged as less accurate than other probe-type technologies(*36*), researchers must take care regarding their assumptions and methodology utilized to analyze network data in order to get credit and meaningful results.

2.3 Pre-Processing Methods of the Network Data

Signals from different cell towers may cover the device of a cell-phone subscriber in any city location. When the signal strength of these cell towers is relatively similar, the signal connection can switch between the cell towers even when the device is stationary(*37*). The user, then, could appear to travel long distances in seconds. This phenomenon is called 'ping-pong' or 'oscillation' in the network(*2*, *9*, *38*). When oscillation happens, the network can witness a fast switch between two or more cell tower antennae, which can be misleading data in trip identification. Because such recorded data cannot be directly used to infer the traveller's movement trajectories, several methods were introduced to mitigate the potential problems of this 'ping-pong' effect (*39*).

2.3.1 Time-Window Based Method

Previous research has identified the time-window based method as one way of mitigating the oscillation phenomenon. To resolve this issue, Diao (*37*) has proposed a time-window based method that eliminates records when the switching speed between two cell towers is abnormal (*25*). In scanning through the record sequences, a short time-window T_w is applied for every record d_0. When the switch speeds between two adjacent cell towers have exceeded the pre-defined time-window T_w, the following record is considered as a 'ping-pong 'effect and will be regarded as a pseudo trip (*38*).

The time-window based method is simple to implement. However, there are several major drawbacks. First, determining the time-window T_w is difficult. If the time window is too short, the oscillation effects may not be removed from the database; on the other hand, if the time window is too long, then true trips will be removed, and the record will contain long-distance travel only. For this method, most researchers determine their threshold based on prior knowledge about one single trip or the characteristics of the data (*9*).

2.3.2 Pattern-Based Method

Since the time-window based method may not solve all oscillations due to the sparse data records, a pattern-based method of eliminating the ‘ping-pong’ effect is utilized by other researchers(*40*). The pattern-based method relies on prior knowledge of the oscillation pattern, in which the signal bouncing between several different cell towers noted as L_0, L_1, and L_2), such as $L_0 – L_1 – L_0$, or $L_0 – L_1 – L_2 – L_0$, where the signal connection will always switch back to the original cell tower(*26*)(*41*). The oscillation sequences are removed after being identified. This method relies heavily on heuristic rules that often struggle to deal with complex situations and that can result in either including pseudo trips or mistakenly removing real travel. The observed pattern is usually connected with the incredible switch speed that is faster than a certain threshold (200 km/h). Sometimes, other types of information are also used to improve the accuracy of identifying real movement, such as trajectories or travel speeds (*3*).

2.4 Trip Identification Methods

After pre-processing, the researcher must identify the moving segment of the records, normally containing one single transport mode. This step is called Trip Identification or segment identification. The purpose of this step is to identify the meaningful stop (the start and end of a single trip) and moving/passing-by records that are realistic trips conducted by subscribers(*19*)(*42*)(*43*). Previous researchers have provided four major ways of conducting transport mode detection from the given data sources. In most the cases, they have attempted to identify the simple transport mode (e.g. subways and light rail transit system) while omitting transport modes that are mixed with other modes such as transit buses, bicycles, or walking (*9*). Due to the temporal and spatial resolution limitation, the mode inference engine normally works when detecting a more general mode group with similar characteristics.

In terms of additional data sources, a geographical road network is commonly used to supplement the main traveller's trajectory data. The geographical information for transport infrastructure can provide additional knowledge to the architecture to better conduct mode detection (*43*).

Georeferencing, as one of the most straight forward and commonly used methods, determines whether the pre-defined geographical coverage of network data intersects with the geographical boundaries of certain train stations, Traffic Analysis Zones (TAZs)(*44*), or municipalities(*45*)(*46*). This method works well in clearly pre-defined areas where the cell towers in both regions are unlikely to have interfered with one another. Georeferencing is commonly used in long-distance travel where people travel from one city to another and has achieved relatively high accuracy rates.

The rule-based heuristic (RBH) method is a set of rules or constraints set up to identify the transport mode between two or several transport modes(*44*). Due to a lack of detailed positioning information, these rules are based on prior knowledge and experience with either the data itself or the characteristics of the transport mode. Travel speed differences between public and private transport are common features that are used to help identify whether passengers are taking the LRT or private cars. Other features like geographical reference, proximity to either road networks or metro lines are also key features that can be utilized to infer the means of transportation.

The frequency-based method (*45*) is deployed when detecting travel at certain times of the day connected with important places (e.g. home, workplace, school, etc.). In previous research, daytime and nighttime are commonly used to identify a traveller's workplace and/or home based on their repetitive travel patterns (*36*). Some studies have also generalized travel trajectories into pre-defined trip types, which can summarize a person's life across a relatively long period (*37*). The frequency-based method is normally employed for long-period analysis, and people with a regular daily routine are more easily identified.

Another method to identify the transport mode is clustering (unsupervised machine learning)(*4*). The k-mean clustering algorithm and hierarchical agglomerative clustering method were both employed in several different papers (*3*)(*5*). These methods classify the 'unlabeled data' into different cluster groups based on their trip characteristics/ features(*17*). This approach selects the first group of data records by measuring the distance between two consecutive points and assesses whether the distance exceeds a certain pre-set threshold. When the distance exceeds the threshold, the time between the two records is examined to see if the traveller is truly moving or not. However, this method is more likely to be used in sighting data where the geographical location of the network data is measured in triangulation using the signal strength between three different cell towers. As such, the method is difficult to be employed for network data with the location estimation using Cell of Origin (CoO)(*9*).

Based on different data sources and data collecting methodologies, these methods all have strengths and weaknesses in terms of identifying the trips conducted. Georeferencing works best in long-distance travel where there are clear, pre-defined boundaries at the start and the end of trips. The rule-based method is good at identifying movement but cannot deal with more complicated scenarios since outliers and missing information can easily skew the results. The frequency-based method can capture trip trajectories fairly well for people with regular daily travel patterns, but a long period of data is required. Last, the clustering method is the most accurate in terms of the geographical location, but the data sources required are limited as the longitude and latitude data for each record are estimated from the triangulation estimation of the signal strength from the nearby three cell stations.

2.5 Transport Mode Identification Method

Due to the spatial and temporal granularity of network data, previous research tries identifying 2 to 3 modes of transport, where train and car are the most popular. Most papers have focused on the inter-city trips where the travel distance is long enough and the Euclidean distance between the train line and highway is greater than the cell station coverage distance(*9*)(*46*)(*47*). For longer travel distance, the recorded data are less likely to be affected by the oscillation phenomenon, and the longer Euclidean distance between different types of transport is easier to detect when considering the relatively low spatial and temporal resolution.

The Rule-Based Heuristic method(*29*)(*48*) is the most commonly used transport mode identification method that involves pre-defined and curated rule sets to identify between different transport modes based on prior knowledge and understanding (*47*). The basic principles are to either compare certain trip features, like the Euclidean distance from the trip trajectories to either rail or road network, or to set up various travel speed thresholds to differentiate between various transport modes. Another way of establishing the ruleset is to construct the geographical reference locations and give the best 'matching' results between several different types of modes (*47*). For instance, the inter-city trip mode detection between two major cities in Canada is deployed by two simple rules where, 1) trip duration between 0.5 to 1.5 hours is assigned as air trip and, 2) trip duration between 2 to 6 hours is assigned as ground transport. Furthermore, more complicated heuristic methods can be implemented in an urban context where trip characteristics of travel speed lower than 8 km/h and travel distance less than 3 km is identified as walking, and those with travel speed greater than 15km/h and no bus stop or train station is within 500 m of the trip destination are classified as car mode. All other trips in between are classified by a logit model whereby travellers decide their transport mode based on the utilities available and their surrounding environment.

In previous studies, cars, trains, and planes are the three modes of transport that are commonly detected, and active transport like walking and biking are difficult to identify due to their low speed. The clustering method(*3*)(*10*)(*11*), which is normally referred to as k-means clustering and hierarchical agglomerative clustering, is another approach to identify transport modes. The main idea is to cluster the selected trip attributes or features summarized from the raw unlabeled data and classify their centroids into multiple different clustering formed around similarities and differences and based on prior knowledge, common sense, or additional information. These methods are powerful when the raw data can provide the triangulated geographical location of the trip trajectories.

The third method is statistical analysis developed to identify transport modes and differentiate between driving, bicycling, and walking. Statistical analysis aims to infer the most likely trip sequence based on the given trips and the historical data of the relevant routes. Xu (*51*) proposed a method consisting of a Hidden Markov Model (HMM) with two different sub-models to identify different transportation modes, where the speed distribution law is implemented to distinguish different travel speeds that are learning from the pre-defined (labeled) training set.

Comparing all three methods, the Rule-Based Heuristic method is the most commonly used method in the research literature, which relies on existing prior knowledge about either the travel behaviour or the means of transport. Common sense knowledge is also implemented, together with the geographical information, which works well for detecting trips along the railways (train, LRT, subway) (*2*). This study presented in this thesis will set rules to classify the transport mode and find the appropriate threshold values between different transport mode clusters. The resulting clusters are interpreted by human analysis due to lack of ground truth, and the final outputs are analyzed based on prior knowledge and intuitive judgments.

CHAPTER 3 METHODOLOGY

This chapter introduces the proposed data processing methods to conduct the transit ridership and OD matrices estimation. The implementation of data pre-processing, trip identification, and the LRT trip probability estimation are also covered in this section.

For the research presented here, the third-party company data provided is fourth generation network data, which has largely improved the temporal and spatial resolution in comparison with traditional CDR and sighting data. The major motivation of this thesis is to address weak or insufficient urban travel dynamic information in other studies by using the new data source to extrapolate more detailed information. The previous research discussed in the literature review here is limited to large-scale Origin-Destination matrices and home and workplace estimations due to the nature of their data types. With a lack of detailed and accurate geographical and time resolution, more detailed travel information like the transport modes and travel routes are elusive. However, with the new generation of the network data and improved temporal and spatial resolution of the data, this thesis will explore the possibility of extrapolating detailed travel information from the data trajectories formed by the network data. The LRT system, which is relatively isolated from road networks and is not impacted by road traffic volume, provides a comparatively simple use case for this method since train travellers should demonstrate similar travel patterns that can be captured by the network data.

To infer the transportation mode, especially in an urban context, the proposed method aims to explore the feature handover behaviour between cell towers along the LRT line. The sequence of feature handover locations can be used to identify whether the traveller is on the LRT line or not since the metro line is relatively isolated from the whole road network. Network data are characterized by spatial-temporal uncertainties that normally require extra data cleaning procedures before performing the detailed modelling analyses. The data utilized in this research are not reported to have an estimation of the spatial location accuracies. In comparison with other research papers with similar data sources, the spatial resolution can vary between 200-400 meters in urban areas where the base stations are denser and the coverage is smaller, and up to several kilometres in suburban areas where the population density is lower.

In terms of the data preparation process, network data requires some data cleaning procedures before conducting any analysis, due to the inherent nature of the data collection process. Previous research has tackled the problem in two respects: temporal and spatial uncertainties (*9*). Depending on the device activities and usage intensity, the recorded data can be heterogeneous and irregular, being more active during normal work hours and less frequent at other times. A common way of dealing with temporal irregularity is to filter out the low-frequency users while focusing on the high-activity users, or else to utilize the data interpolation method to fill the gap between two records with long-time discrepancies.

For spatial uncertainties, the oscillation phenomenon is a common problem that causes noises and pseudo movements in the cell network system. This phenomenon is commonly impacted by the signal strength and load balancing policies in the network. As previously outlined, three types of methods are proposed from previous studies to resolve the problem:

time-window-based, pattern-based, and hybrid methods (the combination both both time-window-based and pattern-based method to improve the accuracy of the transport mode identification). A time-window-based method removes the abnormal records when the device is switching from one base station to another under a certain time threshold. The pattern-based method extracts the record sequence that matches with certain pre-defined switching patterns between base stations (e.g. $L_0 – L_1 – L_0$, which L_0, L_1 represents the base station) and labels the extracted record sequences as oscillation. A hybrid method, on the other hand, combines the two methods to identify more complicated scenarios while reducing the elimination real trips.

In this study, the network data was pre-processed using a hybrid method to extrapolate the possible travel sequences from the raw network data, which considered both trajectory pattern match and travel time when identifying transit trips. The newly proposed hybrid method is different from previous research that served a more dedicated purpose to filtering all the non-LRT travellers that are currently not within the boundaries of LRT line or remaining idle at the time. The newly proposed method is able to take better advantage of the finer time intervals between pairs of records that are passively generated during the travel movements of subscribers.

In previous studies, both temporal and spatial problems were considered when conducting the data cleaning procedure (*2*)(*9*). In the present study, travel patterns have been defined as the cell tower clusters that can stably cover the LRT stations and routes between stations. All cell towers covering a specific station have been denoted as the station cluster, while the cell towers that only cover the path between stations are denoted as the path cluster. The handover behaviour between clusters are the handover pattern used to identify whether the travellers are riding the LRT or not.

The proposed method here can be regarded as a hybrid heuristic method where the algorithm combines the pattern-based and time-window based method to improve the accuracy of mode identification. Travellers not only have to meet the sequential pattern constraints along the LRT but also match the travel time constraints of the LRT train. The overall workflow below indicates the general process of identifying the candidate trip trajectories and the selection process for each step the data. The first step is the pre-processing, where the pattern-match based method and the time-window based methods are both implemented to select the travellers with the potential to take the LRT as part of their journey. The overall workflow of the proposed method is shown in FIGURE 3.1:

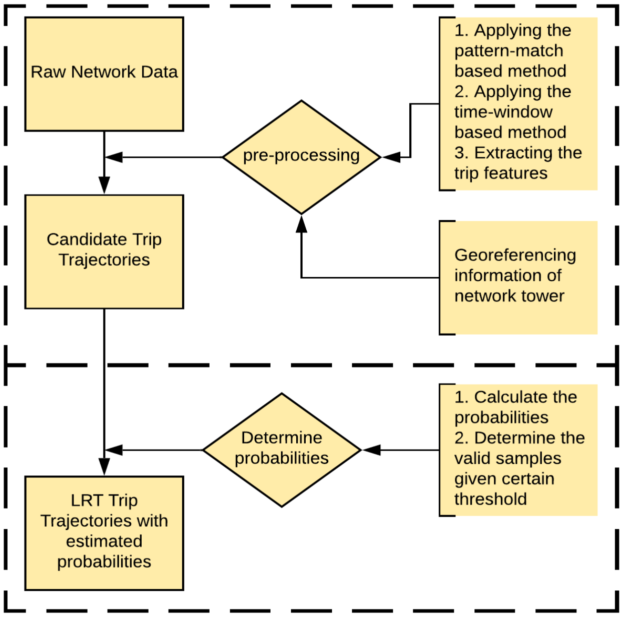

Raw Network Data
pre-processing
1. Applying the pattern-match based method
2. Applying the time-window based method
3. Extracting the trip features
Candidate Trip Trajectories
Georeferencing information of network tower
Determine probabilities
1. Calculate the probabilities
2. Determine the valid samples given certain threshold
LRT Trip Trajectories with estimated probabilities

FIGURE 3.1 The overall workflow of the data processing architecture

3.1 Data Preparation

The main data in this study is the network data that recorded billions of rows representing various activities and location information collected from users. In this research, the network data provided includes approximately 30% residents and travellers out of 1.27 million total population in Edmonton metropolitan area, Alberta, Canada. The original dataset consists of two consecutive weeks in September 2017 that covered both residents and travellers within the Edmonton area at the time. The network data records were produced not only when subscribers called, messaged, and data browsed, but also collected 'ping 'information that was triggered by the network itself to examine the status of the cell phone. The active 'ping 'records decreases the time interval between two consecutive records from several hours to several minutes, greatly increasing the temporal resolution of the data. All the records were passively generated, and the record frequency was still dependent on cell phone usage. The spatial and temporal characteristics are presented for the network data, and the scale of the data and geographical scope of the study are explained in the following section.

The study region covers the only LRT line, the Capital Line, that is operating in a North-South direction consisting of 14 stations. The LRT line travels from residential suburbs in the North-East quadrant of the city, through the Downtown area, the new Rogers Arena, the University of Alberta's North and South Campuses, and several major transit hubs, which makes it an important transport corridor. In this research, the detailed LRT line information was downloaded from the Edmonton Open Portal that is independent of the road network. As shown in **FIGURE 3.2**, the LRT line can be extracted into a set of nodes representing the stations and a set of links representing the rail lines.

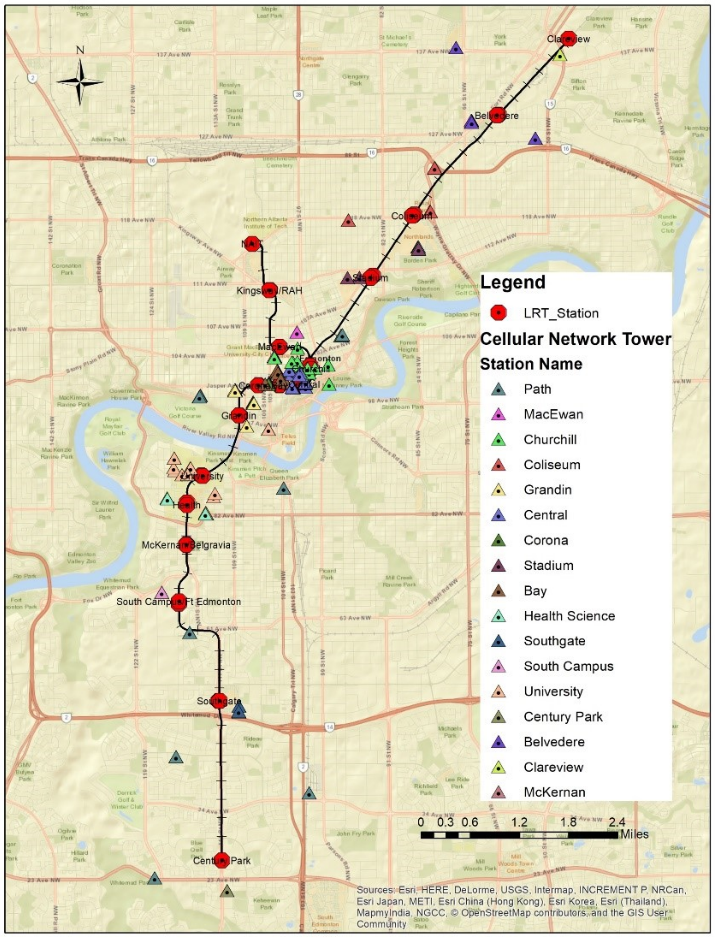


FIGURE 3.2 The study area includes all the stations on the Capital Line and its corresponding base stations.

3.2 LRT Cell Tower Clusters

The information collected from field tests is plotted in FIGURE 3.2. The cell tower clusters are classified into three categories: on-surface station, underground station, and on-route path. Because there is no underground base station, subscribers can only get a connection with the network when he or she leaves the underground area and reconnects with the surface base station. Subscribers who travel along surface stations and on-route paths can normally connect with the network. Therefore, their travel trajectories can be captured during the whole trip.

As seen in FIGURE 3.2, the adjacent base stations have been labelled with their covered LRT stations showing in different colours. The geographical coverage is based on the field tests conducted for eight times in four different days in March 2019. The field test data collects all the base station information that is adjacent to the LRT stations and connection opportunities for passengers.

The data was generated once per second, providing the most comprehensive information possible regarding cell tower connection in real-time. For this study, there was a buffer zone created around each station, in which all GPS points that fall within the buffer zone were considered in-station GPS points and all other GPS points were considered on-route GPS points.

The cell tower information embedded in the GPS records was then labelled and grouped as station and on-route towers to help define the possible travel pattern when people are taking the LRT. Because one traveller could potentially connect with several cell towers and antennae, their record pattern along the LRT line would be the combination of various sequential potential cell towers that cover different stations.

3.3 Defining the Travel Pattern and Train Trajectories

At any given location, travellers could connect with several signals from different cell tower antennae with similar signal strength; the network determines the connection depending on many factors, like volume load on each antenna. The network connection bouncing between several antennae is the oscillation in the network. Therefore, to accommodate these various, potential occurrences when traveling on the train, the proposed handover pattern categorized the cells into several nodes and links representing the LRT line.

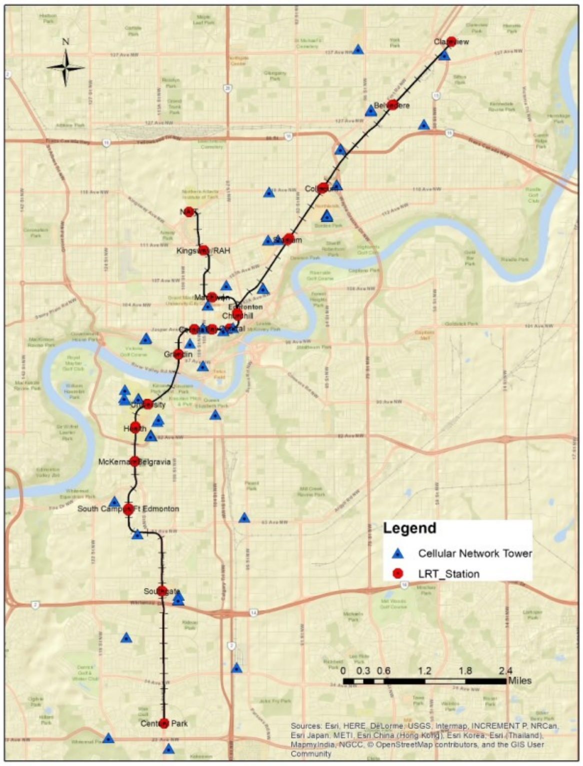


FIGURE 3.3 demonstrates the LRT stations and the geographical locations of the surrounding base stations.

This step is equivalent to a transport land-use partitioning that correlates the LRT line segment by segment with the mobile network. According to the field-collected data and the given theoretical coverage of base stations, labels were given to each station and link to represent the probability of users taking the LRT.

3.4 Pre-processing Algorithm: Heuristic Hybrid Identification Model

A hybrid method was introduced in this research to overcome the signal oscillation and to identify the real movements of subscribers that happens between adjacent cell towers. The method incorporates two criteria that combine both speed requirements and pattern requirements to allow the researcher to infer whether the person is travelling along the LRT line. For one subscriber who connects with one of the potential pre-defined LRT clusters, the algorithm will search for the next record that falls within other clusters within the time threshold. For instance, Subscriber A has a sequence of network data records as follows:

ID	Timestamp (s)	Matched Records	Clusters
178********	2609	AB11191	Century Park
178********	2632	AB11191	Century Park
178********	3197	AB16051	Southgate

TABLE 3.1 One example, Subscriber A, with a sequence of network data

As shown above, the ID is the unique identification code each subscriber possesses, which will remain the same, even on different days, to trace down the network activities. The Timestamp, on the other hand, uses the UNIX timestamp to demonstrate the time when the network system collects the record passively. The latter two columns are defined in accordance with the research scope. The Matching Records show the cell to which the subscriber is currently connected and that falls within the defined LRT geographical area. The Cluster represents the station to which the cell is belongs.

The pre-processing algorithm selects the candidate trips that fulfil both speed and cluster pattern constraints. In this section, the signal oscillation and location uncertainties are detected by both the time-constraint and pattern-based methods, which largely eliminates the risk of identifying the pseudo-trips in the process.

The pattern-matching method is different from the previous studies discussed since the signal coverage of cell towers is assumed not to overlap. In the real world, the user can connect with multiple surrounding base stations simultaneously, and the connection is determined either by the signal strength at the time or station load balancing policies. Thus, the oscillation phenomenon and location uncertainties would be more prevalent in the 4G network since the temporal resolution is much higher. Traditional pattern-based methods have difficulty dealing with the increased number of records, and a simple ($L_0 – L_1 – L_0$, or $L_0 – L_1 –L_1 – L_0$) pre-defined pattern could result in many pseudo trips. In this study, a novel approach to better conduct the pattern-matching algorithm is proposed.

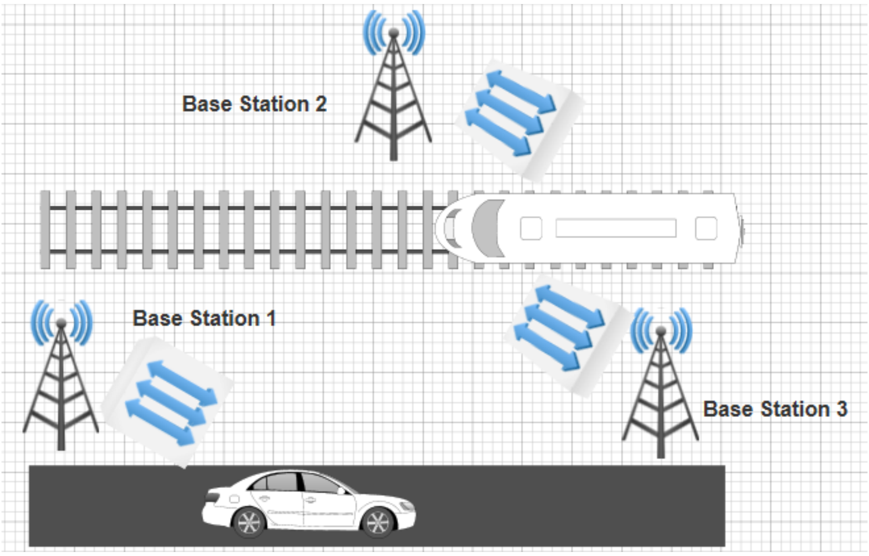


FIGURE 3.4 The base station coverage demonstration

As shown above, base stations will cover different parts of the network, from which the passengers riding the LRT train can receive the signal communication sequentially. The generated travel trajectories formed by a series of network data records are important for identifying the transport mode since repetitive travel patterns should indicate everyday travellers. As the travellers ride the LRT train, their passively generated network data should align with the base stations that are covering the LRT line. As shown in the FIGURE 3.6, the LRT passengers should receive the signal from base station 3, to base station 2, to base station 1, sequentially since the moving train is entering and leaving each base stations' coverage area one by one.

The time-constraint method is drawn from previous research (*5*)(*9*), in which a small time-window $T_{threshold}$ is introduced when scanning through the record trajectories. From previous studies (*3*)(*5*), the 15-minute time threshold is considered reasonable to filter out most of the oscillation phenomena in the database. As shown in the descriptive statistics of the network data, 5 minutes is the average update interval of the network data, which, for more than 80% of the users, averages at least 4 records per hour. The 15-minute time threshold was selected to cover most of the users without being too aggressive with the data filtering and inadvertently losing important travel information.

Due to the 4G network's improved temporal resolution, this study found the $T_{threshold}$ sufficient to filter idle users and travellers who deviated from the LRT line. A user was considered as idle under several conditions; 1) the geolocation reference stopped updating for time longer than the $T_{threshold}$, or 2) The geolocation reference updated but, according to the algorithm, the movement was considered as oscillation phenomenon, or 3) the user did not receive any record updating events and/or the record ceased updating for a time longer than the $T_{threshold}$.

Algorithm 1. Pre-processing

```
Input: Coordinates of Base Transceiver Stations;
Pre-defined cluster pattern groups for LRT line;
Raw input network data;
Output: Matched candidate sequential trajectories;
For each User ID j do
|   search for the record sequentially that is matched with the cluster pattern groups;
|   if record i ∉ pre-defined cluster pattern groups;
|   |   i = i + 1;
|   else
|   |   if record i ∈ pre-defined cluster pattern groups;
|   |   |   add record i in new_trajectory;
|   |   |   i = i + 1;
|   |   else if $t_{i+1} – t_i < t_{threshold}$;
|   |   |   add record i+1 in new_trajectory;
|   |   else
|   |   |   the new_trajectory is finished and stored;
|   |   end
|   end
end
```

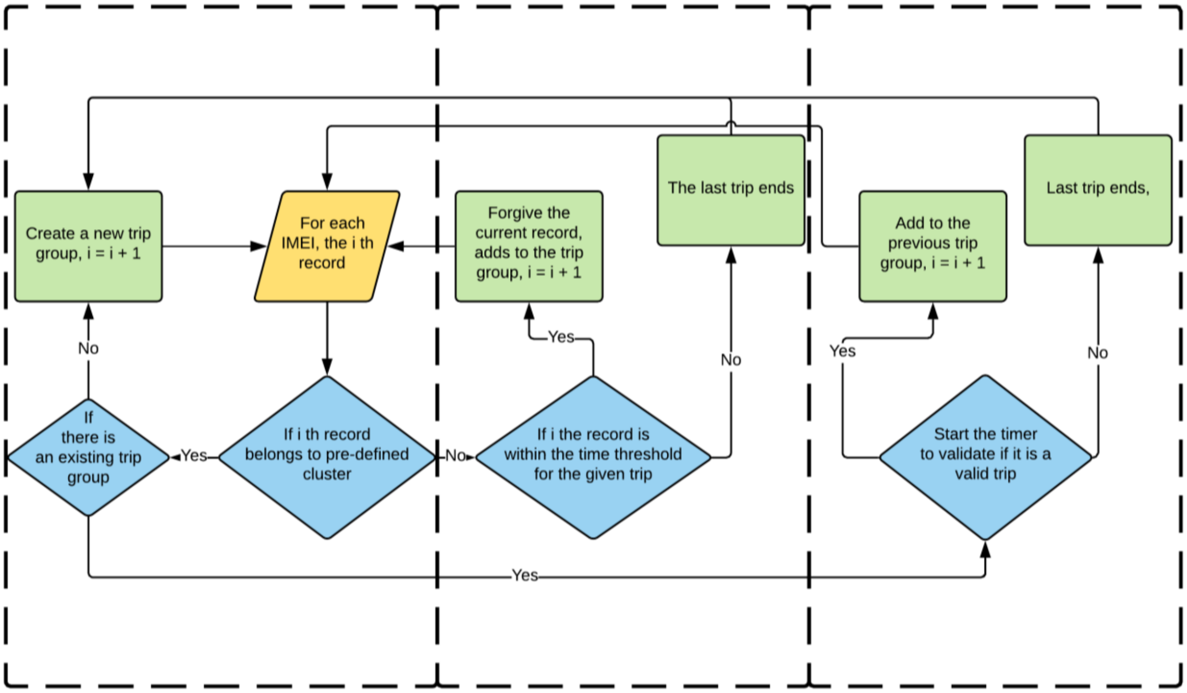

Create a new trip group, i = i + 1
For each IMEI, the i th record
Forgive the current record, adds to the trip group, i = i + 1
The last trip ends
Add to the previous trip group, i = i + 1
Last trip ends,
No
Yes
No
Yes
No
If there is an existing trip group
Yes
If i th record belongs to pre-defined cluster
No
If i the record is within the time threshold for the given trip
Start the timer to validate if it is a valid trip
Yes

FIGURE 3.5 Workflow for trip identification

The algorithm in FIGURE 3.5 demonstrates how the raw network data is processed and grouped into travel trajectories with decision programming. The trip identification selects the candidate trip clusters travelling along the LRT railways that could potentially be LRT trips. All other trips that either do not travel along the geographical location of the LRT railways or possess a staying time that is longer than the threshold value is considered other travel behaviour and will not be evaluated in the following model. The candidate trip clusters are trips that match with both pre-defined trip pattern and the time constraints. The travellers are considered to be the passengers of either the LRT line or private vehicles along the road network besides the LRT railway.

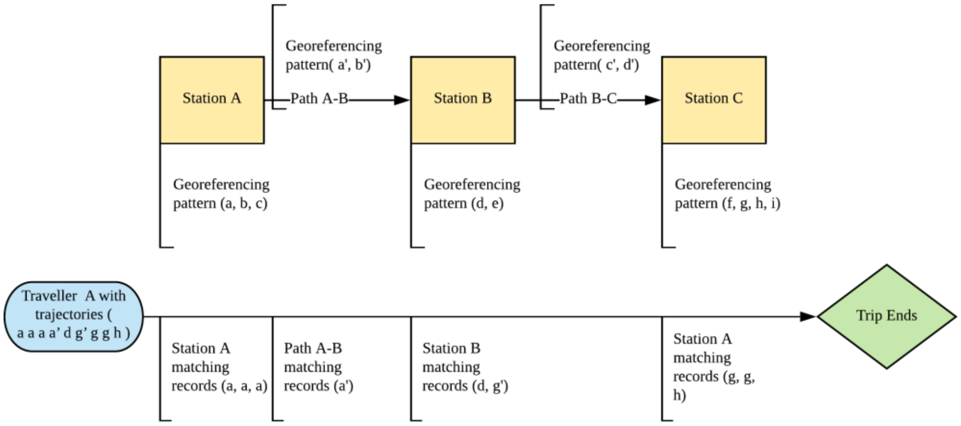

Georeferencing pattern(a', b')
Path A-B
Georeferencing pattern(c', d')
Path B-C
Station A
Station B
Station C
Georeferencing pattern (a, b, c)
Georeferencing pattern (d, e)
Georeferencing pattern (f, g, h, i)
Traveller A with trajectories (a a a a’ d g’ g g h)
Trip Ends
Station A matching records (a, a, a)
Path A-B matching records (a')
Station B matching records (d, g')
Station A matching records (g, g, h)

FIGURE 3.6 One example of pattern matching for identifying an LRT trip

FIGURE 3.6 shows Traveller A with trajectories {a a a a ’d g ’g g h (each letter represents a location identifier)}. The matching pattern given from Station A to Station C derives from the stable cells that are covering the station or the path between two stations. Due to the self-organized algorithm in the cell system network, travellers could potentially connect with any of these stable cells. Some noise can be detected as oscillation happens; however, these stable cells give a clear indication of the location information. Due to sparse records and relatively low temporal resolution in comparison with GPS data, there could be either stations or paths that are omitted during the trip. For instance, Traveller A has skipped the georeferencing pattern Path B-C. However, this does not affect the whole trip being intact as the pattern matching process would only stop when either Traveller A left the LRT geographical area completely or remained at one location for a longer period of time.

3.5 Binary Logistic Regression Model

According to the temporal and spatial constraints from the last step, all travellers who are moving along the LRT line are considered candidates for further analysis. In the pre-processing algorithm, the hybrid time-constraint and pattern-based method only selects potential LRT travellers but cannot accurately identify whether the person is actually on the LRT or in another vehicle. All prior studies have stopped at this stage as either their data sources do not have enough time or spatial resolution to support the analysis, or else they only define the geographical meaning of their network data separately. Yet, the travel trajectory formed by a sequence of network data could potentially provide more information about how the subscribers travelling around the city. Therefore, in this research, the travel trajectories are important since their relative positioning of each cell shows how the trips are made.

The binary logistic regression model is proposed to capture the travel features one LRT passenger could have. The features are constructed intuitively to represent the possible network data trajectories when subscribers are considering taking the LRT as their mode of transport.

3.5.1 Feature construction with the transport networks

A travel trajectory is a sequence of visited inferred network locations utilized in this study to exam the probability of taking the LRT during the trip. Here, the selected features from the visited inferred network location sequence were constructed based on the related information between the network tower and the transport network. To construct the travel features, prior knowledge and intuitive determination were employed to select the best context framework for considering the probability of taking the LRT during the trip. Major arterial roads, tramways, and train stations were downloaded from the Edmonton Open Portal, which provided the geographical information about the selected features. All the following features are constructed to represent the trip characteristics:

- $N_{station}$: Number of station clusters connected during one trip
- N_{route}: Number of on-route path clusters connected during one trip
- N_{match}: Number of records where travellers are matched with the pre-defined pattern
- P_{match}: Percentage of records where travellers are matched with the pre-defined pattern
- T_{match}: Time length that travellers are connected with the pre-defined network clusters
- $P_{t\text{-}match}$: Percentage of time during the whole trip that travellers are connected with the pre-defined network clusters
- $T_{difference} = \frac{|T_{total} - T_{expect}|}{T_{expect}}$
    - T_{total} is the total travel time for the current trip from the origin station to the destination station
    - T_{expect} is the expected travel time for the same origin station to the destination station

To more accurately interpret the candidate users from the network data, several features were extrapolated from the raw data to help to identify LRT passengers. Based on prior knowledge about LRT operation and passengers' cell phone usage habits, several assumptions were made to more accurately infer the probability of taking the LRT as transport when travelling. First, the LRT is a specialty form of transit that operates on a fixed rail and therefore has a determined route, unlike private vehicles and bikes that allow passengers to vary their routes by choice. Therefore, the longer the passenger travels along the LRT, and thus the determined, mapped route, the higher the probability that the passenger is taking the LRT train. Second, the more network records and connections made with the cell towers belonging to the pre-defined clusters, the higher the probability that the candidate passenger is taking the LRT train. In these field tests, all the collected cell tower information is from relatively stable cell towers when taking the LRT; people who are connected with these cell towers have a higher probability of taking the LRT at the time. Third, since the LRT trains are regularly on schedule and without traffic interruptions, the travel time for every passenger should be relatively stable from boarding to alighting. Thus, those who travel with abnormal speed (i.e. too fast or too slow from the expected travel time) are considered less likely to be on an LRT trip. Based on these assumptions, there were nine features extracted from the pre-processed results as input for the binary logistic model to identify the probability of the candidate passenger taking an LRT trip.

The probability $\pi_{i,lrt}$ of one candidate Traveller i taking the LRT is defined as:

$$\pi_{i,lrt} = logit^{-1}(u_i) = \frac{e^{u_i}}{1+e^{u_i}} \qquad (3.1)$$

where all other means of transport have been classified as ‘other’ in this model. The logit u_i follows the linear model that has been built from the training dataset, where Traveller i with a probability of taking LRT $\pi_{i,lrt}$ higher than the probability of taking other transport modes $\pi_{i,other}$.

$$y_i = \begin{cases} 1, \pi_{i,lrt} > \pi_{i,other} \\ 0, otherwise \end{cases} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.2)$$

The model is designed to select these LRT trips from all other trips undertaken by subscribers. Therefore, a higher probability indicates that the trajectory features are closer to an LRT trip as understood from the prior knowledge and intuitive understanding. Lower probability indicates that the person is distant from the LRT line and is travelling with no comparable pattern to an LRT trip.

Variables in the Equation

Step Number	Variables	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	$N_{station}$	1.287	0.271	22.482	1	0.000	3.621
	N_{route}	0.564	0.310	3.299	1	0.069	1.758
	P_{match}	0.047	0.029	2.535	1	0.111	1.048
	N_{match}	4.625	1.092	17.924	1	0.000	101.967

	$T_{difference}$	-5.474	0.938	34.067	1	0.000	0.004
	T_{match}	0.000	0.001	0.002	1	0.966	1.000
	$P_{t\text{-}match}$	0.142	0.122	1.345	1	0.246	1.152
	Constant	-7.102	1.433	24.567	1	0.000	0.001
Step 2	$N_{station}$	1.288	0.270	22.665	1	0.000	3.624
	N_{route}	0.560	0.298	3.540	1	0.060	1.751
	P_{match}	0.047	0.029	2.547	1	0.111	1.048
	N_{match}	4.616	1.071	18.572	1	0.000	101.053
	$T_{difference}$	-5.478	0.933	34.441	1	0.000	0.004
	$P_{t\text{-}match}$	0.141	0.122	1.343	1	0.247	1.152
	Constant	-7.077	1.309	29.231	1	0.000	0.001
Step 3	$N_{station}$	1.317	0.272	23.432	1	0.000	3.730
	N_{route}	0.655	0.287	5.196	1	0.023	1.926
	P_{match}	0.042	0.029	2.113	1	0.146	1.043
	N_{match}	4.491	1.069	17.659	1	0.000	89.222
	$T_{difference}$	-5.511	0.933	34.880	1	0.000	0.004

	Constant	-6.686	1.254	28.409	1	0.000	0.001
Step 4	$N_{station}$	1.5.3	0.248	36.807	1	0.000	4.497
	N_{route}	0.766	0.277	7.653	1	0.006	2.151
	N_{match}	4.396	1.046	17.656	1	0.000	81.166
	$T_{difference}$	-5.322	0.911	34.157	1	0.000	0.005
	Constant	-6.769	1.231	30.235	1	0.000	0.001

TABLE 3.2 The seven features extracted from the pre-processed results and the selected features that have a significant impact on determining the candidate trajectories

According to previous research (*2*), prior knowledge about LRT operation, and intuitive understanding, the following seven features are classified for each trip trajectory for future analysis: 1) Number of stations connected, 2) Number of on-route path-connected, 3) Number of records that match travellers with the pre-defined pattern, 4) Percentage of records that match travellers with the pre-defined pattern, 5) Time length that travellers are connected with the location clusters, 6) Percentage of the travel time that travellers are connected with the location clusters, and 7) Travel time difference.

Concerning the first feature, it is assumed that the longer the person is travelling along the LRT route, the higher the likelihood the person is taking the LRT train. Thus, the number of stations to which a particular trip is connected can be used as an indicator for detecting the transport mode. For the same reason, the number of on-route paths (the second feature) can also be utilized to indicate whether the person is travelling along the LRT line or is geographically proximate to the LRT. Another feature is the number of records that match travellers with the pre-defined pattern. In addition, when the traveller is moving along the LRT line while using the cellphone, the corresponding records should be connected with the adjacent base stations. Despite possible outliers where the cellphone is connected with one base station further away, travellers should still commonly have a high percentage of records that are connected with the adjacent base stations, especially during the travel time when the towers switch between different base stations. Another feature is the total time that the cellphone is connected with the adjacent base stations. It is assumed that the longer the time the cellphone is connected with the surrounding antenna, the higher the possibility that the person is taking the LRT. The last feature is the travel time difference. The LRT is rail-based and has signal priority when travelling through the city. This would suggest that the travel time is not likely to be impacted by on-road, vehicular or pedestrian traffic. In that case, the travel time between each station can be estimated confidently according to the schedule. Therefore, the travel time difference between the traveller and the theoretical travel time estimated from the LRT schedule is a valid feature to determine LRT ridership.

The backward stepwise method is implemented to determine the best variables to describe the features, and the number of stations matched, number of paths matched, number of clusters matched, and expected travel time error were selected. TABLE 3.2 demonstrates the nine features extracted from the pre-processed results, from which the trip features were used to estimate the probability of passengers taking the LRT. Due to lack of ground truth, 550 candidate trips were manually marked as LRT and non-LRT trips for building the binary logistic model. These candidate trips were determined by the empirical judgement that trips accorded with the LRT speed and the pre-defined travel patterns and was based on the intuitive and prior knowledge about how the LRT line is operating across the city. For instance, we know that the signal disconnects when the LRT is traveling underground in the downtown and university areas; we also know that LRT travel speeds are stable in comparison to private vehicles whose speeds can vary more widely and more frequently.

All these trips were marked to be either LRT or non-LRT trips and the backward stepwise method helped to select the significant travel features to extinguish between the two trips. From the prior knowledge and intuitive judgement, the lengths of trips that matched with the travel pattern along the LRT line, and the speeds of the travellers, helped the model to identify the candidate's trips considered to be taken by LRT passengers.

3.6 Data Validation and Training

According to previous studies (*9*), validating the results generated from the network data with third-party data is difficult. GPS records, one of the most accurate technologies to determine the geolocation of a person carrying a device with GPS tracking technology, will not always be as prevalent or have as high a market penetration rate as with network data. Other traditional ways of conducting travel pattern analysis are based on groups of people. Due to its small sample size, it can only be used to validate general travel patterns like total trip generation and trip attractions. These aggregated results can represent the general trend of the data but cannot represent personal travel trips. Therefore, in this study, some of the standards are defined to estimate the accuracy and reasonableness of the results based on the prior knowledge and intuitive understanding of how people normally travel within the city.

3.6.1 Trajectory Probabilities of Binary Logit Model

The probability density function for a binary logit model is used to identify the homogeneousness of the result distribution after all the analysis. By using the binary logit model, the derived transport mode probability distribution of trajectories is one of the performance indices to help understand how results react to the model.

CHAPTER 4 RESULTS AND ANALYSIS

This section presents the main results generated from the model, the interpretation of the outcomes, the analysis of the current model, and future research possibilities for improvements to the use of big data technologies to infer transport mode.

In recent years, new data sources have provided more accurate and reliable information about how people move across the city, and new technologies come with challenges for researchers to explore. The 3G network data that this research is implemented have been through multiple studies and analysis (*3*)(*5*)(*9*), however, the 4G network data with better temporal and spatial resolution, gives the opportunities to supplement the limitations of traditional household travel survey. The high market penetration rate and large sample sizes are asset that traditional data collecting methods cannot compete. With the improved data quality provided, the LRT transit system in the City of Edmonton are being selected as the study scope as the current transportation planning methods are difficult to monitor the operation level and travel demand. Therefore, this study is focusing on utilizing the network data to fill the gaps and provide a sustainable way to monitor the public transit system.

Transportation mode analysis assesses the mode split between one origin-destination pair and how many people are taking public transit instead of private vehicles. The major identifiers are the travel time and the pattern matching results. As the road test is conducted for the LRT line in Edmonton, the trip pattern for the LRT line is confirmed. To estimate the public transit share, the same method is used to create a trip pattern for the LRT line and bus transit. However, bus transit is difficult to distinguish from general traffic, and more road tests are required to create accurate trip patterns. The road test shows the trip pattern collected for the LRT line. As for the trip pattern matching process, all travellers with a high match rate to the LRT line are considered LRT passengers, where their cell sequences comply with the trip pattern along the LRT line. With the support of the road test data to generate the trip pattern, Edmonton passengers taking LRT trips can be distinguished by a travel pattern match to the road test data.

4.1 Study Area

In this study, the LRT Capital Line was selected as the study area that travels in a north-south direction. The Capital Line is approximately 22 kilometres in length and includes 14 stations in total, traversing the most populated areas in the City of Edmonton. The LRT line passes through several major districts including the University of Alberta's North Campus, the downtown business district, Century Park (southern major transit hub), Southgate (southern major business district), and Clareview (northern major transit hub). Century Park and Clareview are also hubs in the Edmonton Park and Ride program. Therefore, the daily passenger boarding, and alighting patterns are representative and can be used to validate the results.

In network data, the location information is based on the cell tower antennae. According to cell tower information provided by the operators, the theoretical coverage of each cell tower and antenna can be calculated. However, different physical environments, different weather, and operation purposes, could affect the real coverage as compared to the theoretical coverage. To mitigate these differences, field tests were conducted along the LRT line that aimed to plot the real connections between cell towers and LRT passengers.

4.2 Field Test Results

In this research, the field test was an important component to acquire the baseline of real cell tower coverage. As explained above, real cell tower coverage can be very different from the given theoretical coverage or normally assumed Voronoi polygon shape. Therefore, to accurately map the handover features and the stable cell cluster covering station and path, field tests were used to map the handovers between cell tower antenna. The GPS records are the best way to verify the individual trajectory result estimation, although GPS data requires additional equipment to collect geographical information. As a result, the data can only be implemented on a small scale compared to network data that covers more than 25% of the total population.

In this study, one complete trip from Century Park station to Clareview station was collected with both GPS records and network data available at the same time. This was also the only trip that had GPS ground truth that could be used as a reference to compare accurate trajectories and our model. As seen in FIGURE 4.1, the model has captured the whole trip of the target trajectories with the underground station missing due to lack of cell network coverage.

The field test was conducted in February 2018 during peak and non-peak hours to decipher the cell towers that travellers may connect to during their LRT rides. An Android application was utilized to collect both the GPS data and information on the cell phone-tower connections. The Android application was based on the GPS sensors and GSM modules of Google Nexus 5 to collect the longitude, latitude, timestamp, and current connecting cell tower information once per second. The collected field data were used as the reference data to demonstrate the signal coverage of cell towers along the LRT line.

As we can see below, the idle smartphone is constantly communicating with the network system while the passively generated network data have plotted out the entire trip when the smartphone is traveling with the LRT train from Century Park station to Clareview station in the northbound direction. In the section below, the station number and name represent the traveller's current location along the LRT line, and the matched records demonstrate that the network subscriber is currently communicating with the cell towers that are covering the selected station.

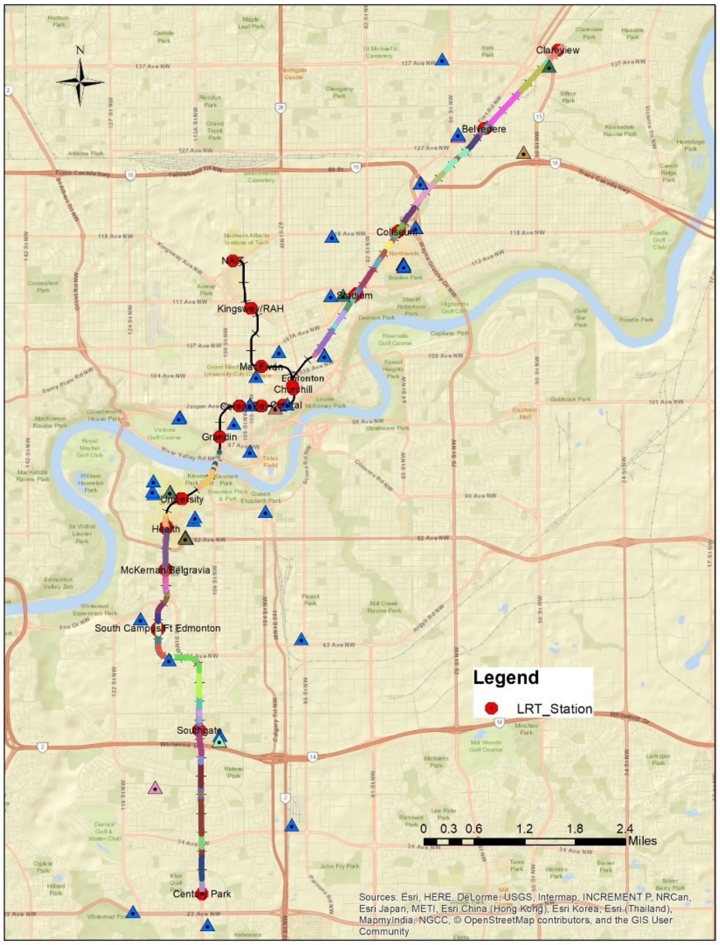

N
Clareview
Belvedere
Coliseum
NAIT
Stadium
Kingsway/RAH
MacEwan
Edmonton
Churchill
Central
Grandin
University
Health
McKernan Belgravia
South Campus Ft Edmonton
Southgate
Century Park
Legend
LRT_Station
0 0.3 0.6 1.2 1.8 2.4
Miles
Sources: Esri, HERE, DeLorme, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), MapmyIndia, NGCC, © OpenStreetMap contributors, and the GIS User Community

FIGURE 4.1 One example of the field test along the Capital Line

Due to missing information from the reference data, there was no third-party dataset that could be utilized for comparison and result validation. In that case, the field test data in this study became an important independent data source that could validate the proposed methodology. As stated in the literature review, the rarity of validation data and methods has been a consistent drawback for network data, adding uncertainty to the analyzed results.

4.3 Descriptive Statistics of Network data

Data used in this research consists of 4G Network data from approximately three hundred thousand users in the Edmonton metropolitan area during two complete weeks in 2017. The database includes all active devices that were connected with the network of mobile carriers within the Edmonton metropolitan area. The network data were collected by mobile operators for operation and billing purposes and to ensure that mobile users remained connected with the network. Each record contains the User ID (encrypted user ID), a timestamp, cell tower information, and type of activities. According to the descriptive statistics, the mobile users included in the records could be long-term residents, short-term residents, visitors and/or commuters.

Before any processing steps, the raw network data contained all three hundred thousand possible travellers within the Edmonton metropolitan area. According to the Edmonton census, the population for the metropolitan areas is approximate 1.27 million. The network data are passively generated by the mobile operators and include all types of users as long as they are active in the network. Unlike traditional methods of conducting the travel survey, network data can capture all types of commuters, including both regular and irregular travel. The raw network data contain about 30% of the total population, and the network data can be viewed as a representative sample in the study.

The network data itself provides some self-selection bias that can skew the analysis slightly and may represent groups that are more active and travel more frequently. The groups with higher cell phone usage will be better captured because of high frequency activity records. The first hour included data from the previous day due to the packing process of the data providers.

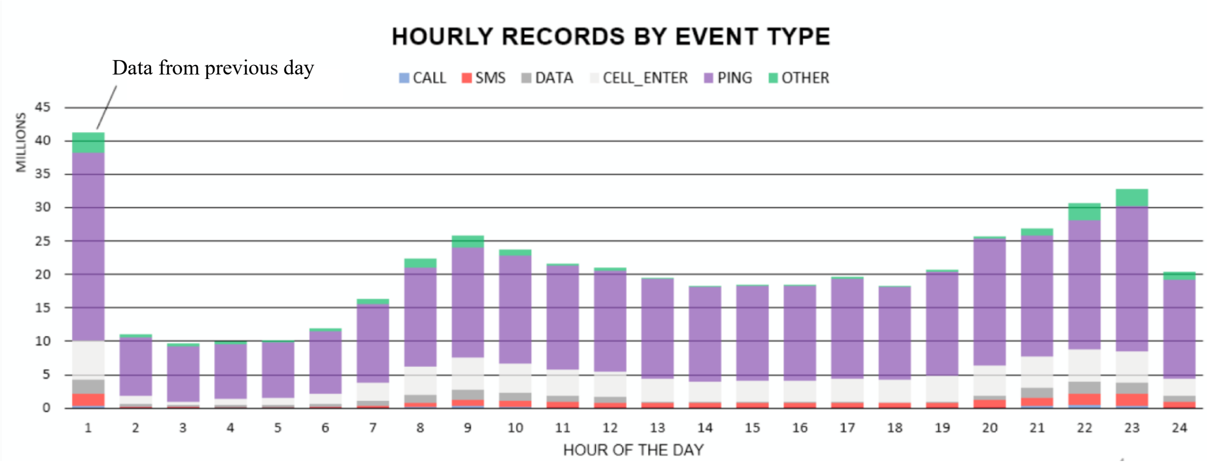


FIGURE 4.2 The hourly recorded network data by event type

In general, the weekly passenger volume pattern for each station is presented below, where the usage of each station on weekdays is seen to be repetitive. The week of data collection included a national statutory holiday on the Monday as shown. The output result also demonstrates that there was a significant drop in volume during both morning and afternoon peak hours. Also evident is that the Century Park Station, University District, and Downtown areas were three hotspots with large amounts of passengers boarding and alighting at these stations. The passenger number estimation for working weekday have shown in great consistency as university and downtown areas have greater travel demand, in contrast to the passenger number in weekday and statutory holiday.

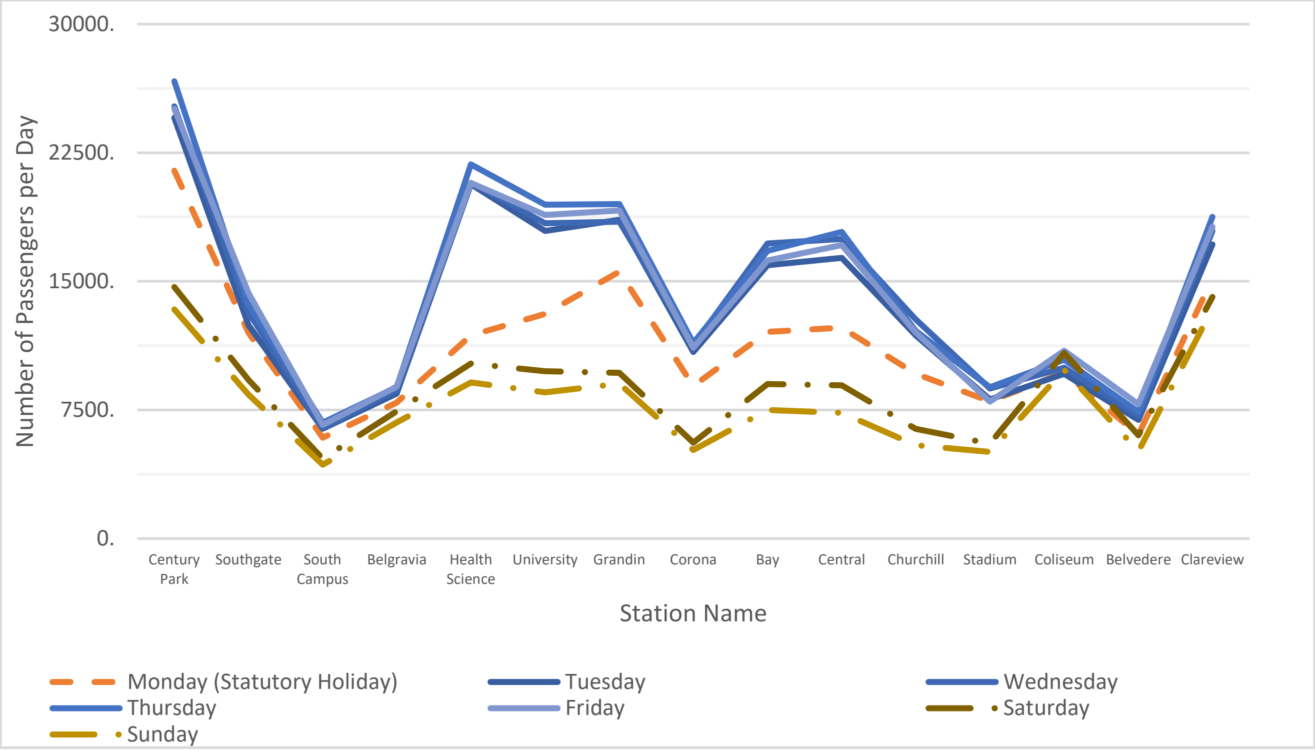


FIGURE 4.3 The weekly pattern of boarding and alighting passenger volume for each LRT Capital Line station

4.4 Performance Index of Transport Mode Inference

4.4.1 Trajectory probabilities

Using the binomial logit model, the probability distribution of trajectories was derived below. The distribution of probability for the different types of transport mode is shown in FIGURE 4.4, which demonstrates how the model estimates the different types of transport mode during the trips. For most of the identified candidate trips are being categorized as non-LRT trips since the trip identification step have included all trips conducted near the LRT line. However, those trips are mostly being considered as other modes of transport in our transport mode inference method.

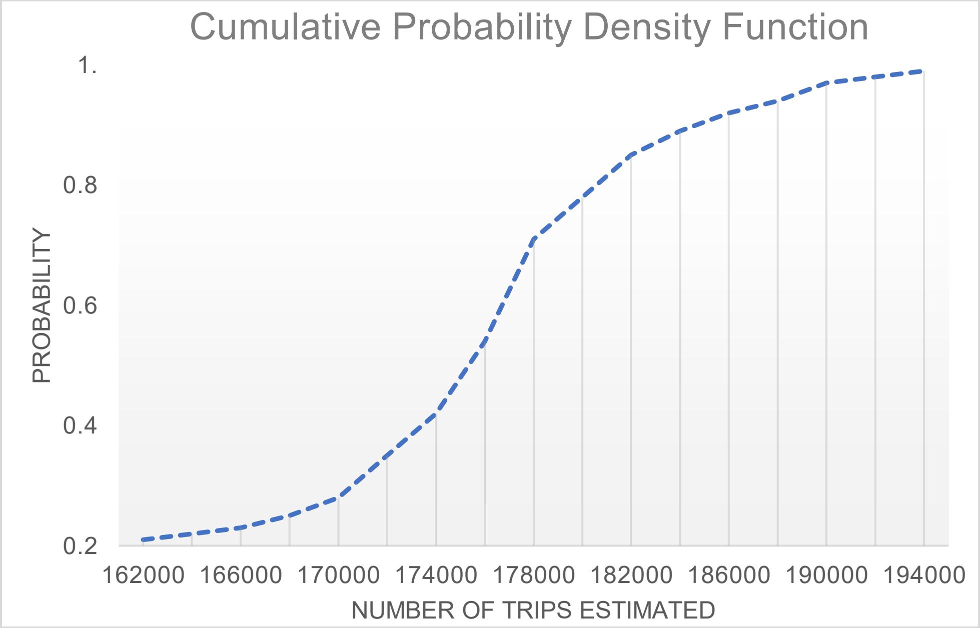


FIGURE 4.4 The probability density function for the binary logit model

4.4.2 Cross-Validation with LRT Annual Report

The results presented are derived from the network data gathered over two weeks in September 2017 and was compared with the LRT annual report 2017 that is collected and analyzed by the City of Edmonton for the general information about the LRT operation and passenger counts for each station. For comparison, the estimated passenger counts were projected to the whole population with a 95% confidence level. The annual report used the traditional methodology of collecting the passenger data by manually counting the boarding and alighting of passengers at the door from the beginning to the end of the service. During both survey periods, there were no big events such as sports games or major outdoor activities. This is significant because special events can potentially create variance in the passenger count and have an impact on the results. As shown below FIGURE 4.5, the average weekday comparison between the estimated passenger count and the annual report passenger count at each station shows similar trends with minor differences at certain stations.

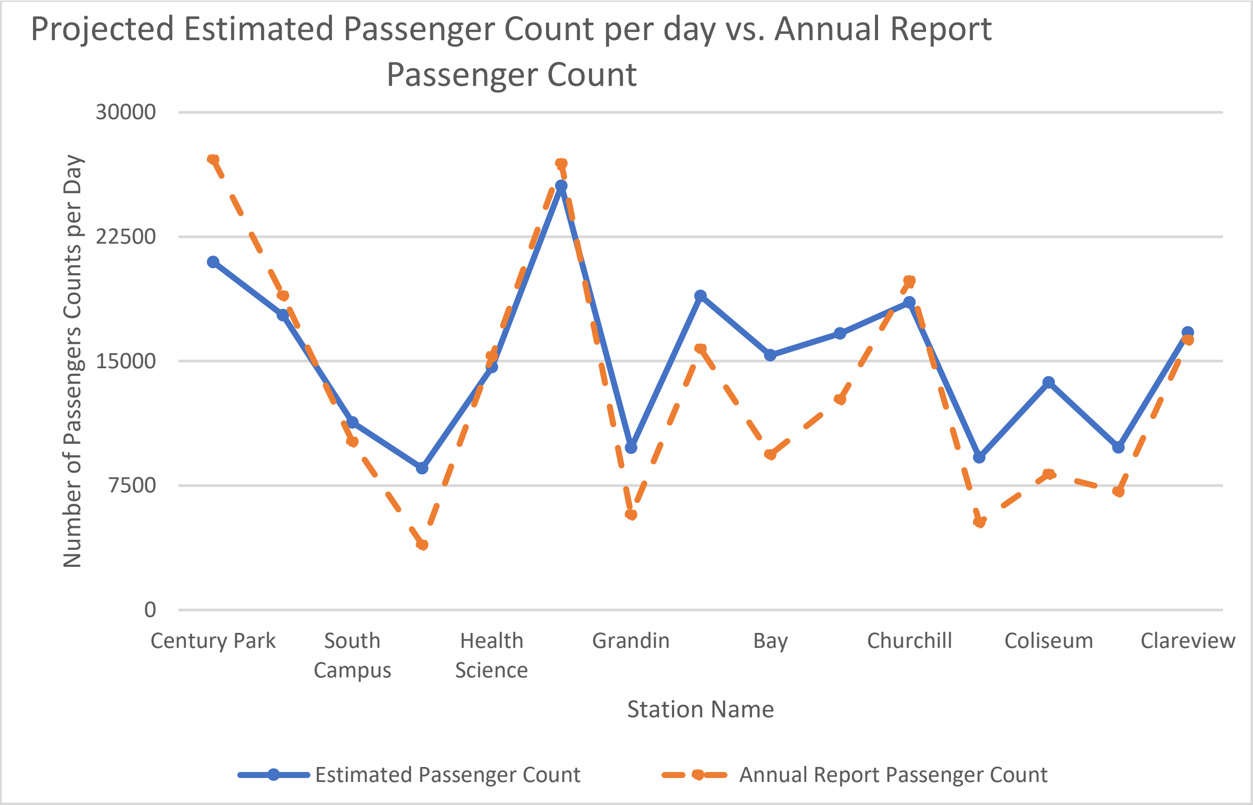


FIGURE 4.5 Projected Estimated Passenger Count per day vs. Annual Report Passenger Count

FIGURE 4.5 shows the daily estimated travellers taking the LRT during the weekday, compared to the real-world morning peak and afternoon peak hours. As per data providers, network data is inclined to have lower update frequency and activity events as people tend to use their cellphones less frequently in the evening. Therefore, the p.m. peak hours demonstrate lower numbers. However, a general trend is still identifiable as more people tended to take LRT between 4 p.m. and 6 p.m. The passenger count increased dramatically, starting at 5 a.m., which is congruent with the first LRT train departing at the same time. Moreover, the passenger count peaked at 7:30 a.m., which can reasonably be assumed to be related to work travel in the morning.

In a general trend, the estimated passenger count is higher than what the LRT annual report is suggesting, which could be resulted in several reasons. For certain stations like Grandin, Corona, Bay Enterprise Square, and Central station where the annual report is significantly lower than what the model is estimated. This may due to that the underground station does not have the cell tower coverage which there is no data being collected while the travellers are underground. The only travel information that the network data can capture is when the travellers are about to entering the station and the first records when the travellers are leaving the station underground area. Therefore, the underground to underground station trip trajectories will be lacking enough travel information, and other trips like private vehicles traveling within the downtown area or pedestrians walking through the downtown pedway system could be miscounted as an LRT trip. As might be expected, the proposed method is more accurate in identifying the long distance trips where there are more collected records along the route. Trips that are conducted within downtown areas where people are simply travelling between different buildings along the LRT line would be tricky to identify.

The Century Park station was the only station where the passenger count from the LRT annual report was significantly higher than the model estimation. One reason could be that the Century Park station is located in the southern part of the city where business activity is low. The cell tower density, as mentioned in the previous section, is significantly lower than that of other core areas like the downtown or university. As such, the subscribers could have connected with other cell towers due to network system balancing procedures. The model, in that case, will not be able to detect if those travelers collected to other cell towers are starting from the Century Park station.

Although the LRT annual report is the only reference document for the comparison analysis, the estimations in the report itself were conducted using a traditional data collection method. The report stated that the passenger count at each station was collected by two data collectors riding the LRT train and counting the number of passengers boarding and alighting at each station. The data collection was conducted over a single day, leading to potential bias based on the small sample size. The difficulty of monitoring actual public transportation use and the current labour-intensive form of data collection provide justification to find better ways to conduct these audits, especially since detailed and accurate information is key for transportation planners, engineers, and decision makers.

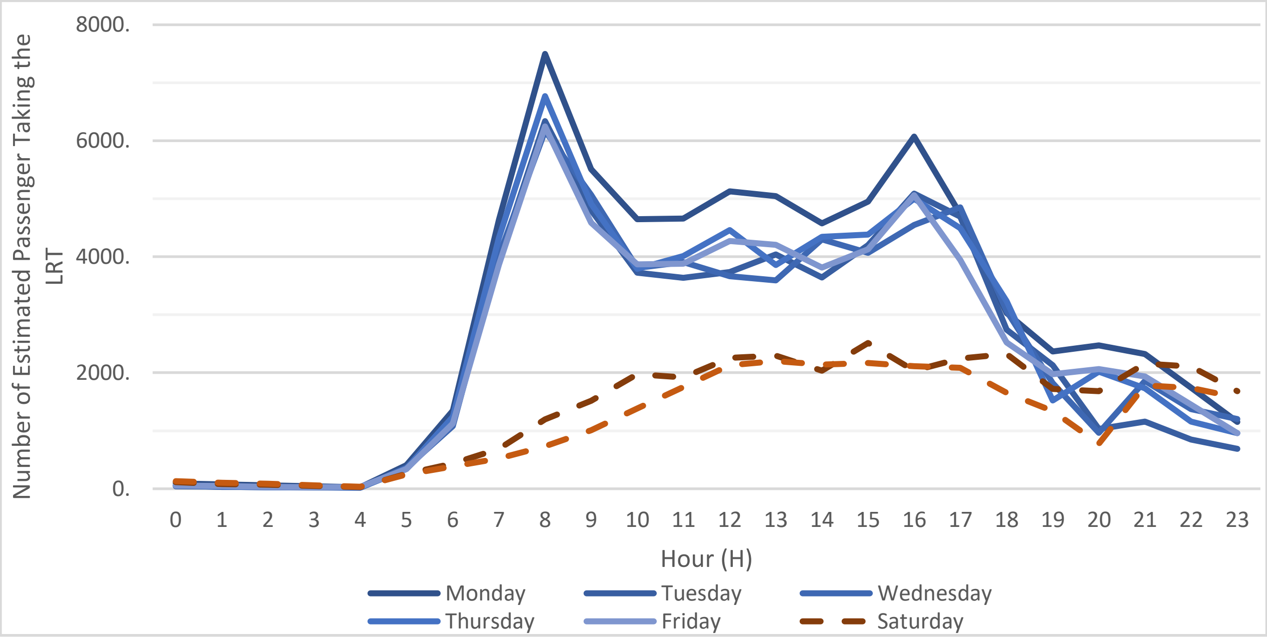


FIGURE 4.6 Daily Estimated Travellers Taking the LRT Capital Line for All Stations

Because the Edmonton LRT annual report does not have an hourly passenger count data, the estimated passenger count was analyzed to match with the prior knowledge that morning and afternoon peak hours should occur during the weekday and the volume of travellers are much higher in weekdays than that in weekends.

Before conducting the pre-processing step, the network data contained all users who commuted through the City of Edmonton. To identify the LRT passengers through the day, the nature of network data was considered. Even though the 4G network increases the temporal resolution of data so that more comprehensive travel trajectories can be plotted, there are still 20 to 30 % of users whose daily number of records are less than 100. In that case, only users whose records were collected by the operator when taking the LRT simultaneously could be identified. Therefore, the first step was to identify candidate passengers whose mobile records matched with the pre-defined LRT line. The candidate passenger trips identified from the pre-processing steps were not the actual LRT trips, but the trips associated with LRT or surrounding road networks.

4.5 Comparison between Proposed Method and Traditional Mode Identification Methods

To record the accurate geographical location of travellers in comparison with the proposed model estimated travel trajectories, this research used built-in GPS sensors in the field test. The following is an example of the road test conducted for validation and comparison for the model on January 16, 2018.

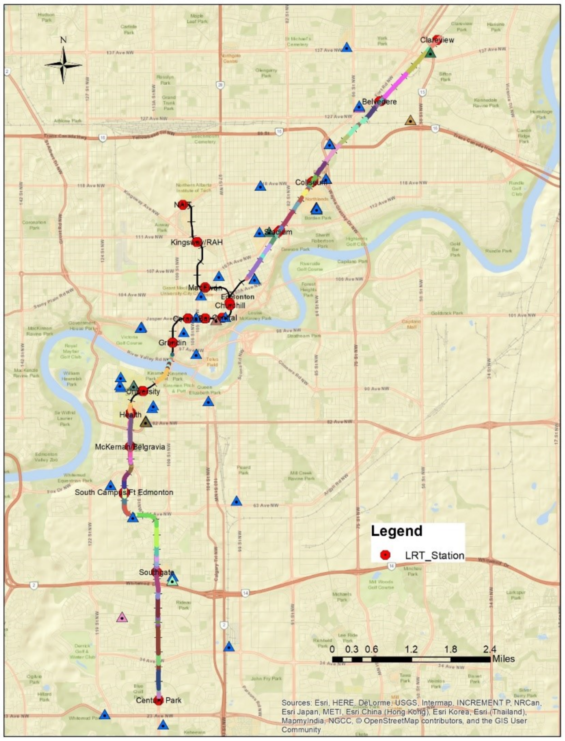
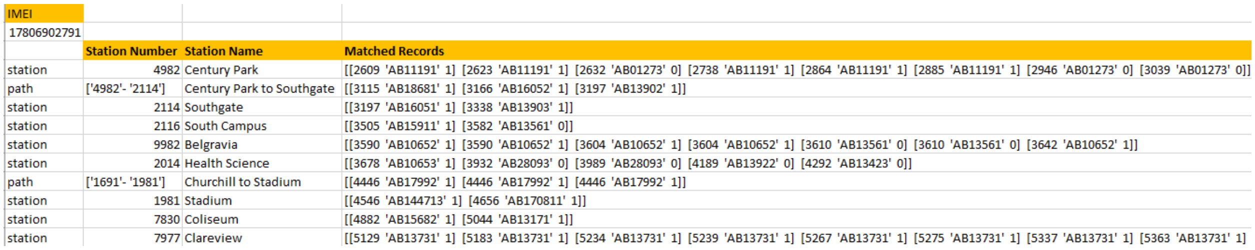

N
Clareview
Belvedere
Coliseum
NAIT
Kingsway/RAH
Stadium
MacEwan
Edmonton
Churchill
Central
Grandin
University
Health
McKernan Belgravia
South Campus Ft Edmonton
Southgate
Century Park
Legend
LRT_Station
0 0.3 0.6 1.2 1.8 2.4
Miles
Sources: Esri, HERE, DeLorme, USGS, Intermap, INCREMENT P, NRCan, Esri Japan, METI, Esri China (Hong Kong), Esri Korea, Esri (Thailand), MapmyIndia, NGCC, © OpenStreetMap contributors, and the GIS User Community

FIGURE 4.7 Smartphone collected GPS records for comparison

As discussed, the test passenger boarded the LRT at Century Park station and alighted at Clareview station, travelling through the whole network. As per the proposed model, the probability of this subscriber taking public transit, based on their long-distance travel route, is likely. This is especially the case because the LRT track is geographically isolated from the major arterial road network. Although the southern part of the track is close to the major roads, the northbound train track is more isolated, and therefore the model can more easily monitor the traveller and identify whether the person is travelling on LRT or other means of transport.



IMEI			
17806902791			
	Station Number	Station Name	Matched Records
station	4982	Century Park	[[2609 'AB11191' 1] [2623 'AB11191' 1] [2632 'AB01273' 0] [2738 'AB11191' 1] [2864 'AB11191' 1] [2885 'AB11191' 1] [2946 'AB01273' 0] [3039 'AB01273' 0]]
path	['4982'- '2114']	Century Park to Southgate	[[3115 'AB18681' 1] [3166 'AB16052' 1] [3197 'AB13902' 1]]
station	2114	Southgate	[[3197 'AB16051' 1] [3338 'AB13903' 1]]
station	2116	South Campus	[[3505 'AB15911' 1] [3582 'AB13561' 0]]
station	9982	Belgravia	[[3590 'AB10652' 1] [3590 'AB10652' 1] [3604 'AB10652' 1] [3604 'AB10652' 1] [3610 'AB13561' 0] [3610 'AB13561' 0] [3642 'AB10652' 1]]
station	2014	Health Science	[[3678 'AB10653' 1] [3932 'AB28093' 0] [3989 'AB28093' 0] [4189 'AB13922' 0] [4292 'AB13423' 0]]
path	['1691'- '1981']	Churchill to Stadium	[[4446 'AB17992' 1] [4446 'AB17992' 1] [4446 'AB17992' 1]]
station	1981	Stadium	[[4546 'AB144713' 1] [4656 'AB170811' 1]]
station	7830	Coliseum	[[4882 'AB15682' 1] [5044 'AB13171' 1]]
station	7977	Clareview	[[5129 'AB13731' 1] [5183 'AB13731' 1] [5234 'AB13731' 1] [5239 'AB13731' 1] [5267 'AB13731' 1] [5275 'AB13731' 1] [5337 'AB13731' 1] [5363 'AB13731' 1]

TABLE 4.1 The stepwise result for the same trip from the network data

As shown in TABLE 4.1, the tester was identified as boarding the LRT line at Century Park station, heading northbound. The surface stations both on the south and north side of the network are easily identified since the network connection was strong and updated frequently during that time. The underground stations that are not covered by the network tower signals are between the Health Science station and the Stadium station. This large signal offline gap can also be utilized as a travel identification feature since other means of transport, like private vehicles and transit buses, are still on the ground; therefore, the network connections are still in progress and handling activities can still proceed.

4.6 Passenger Statistics and Origin Destination Estimations between Stations

A major result generated from the proposed model to represent estimated transit ridership from the network data is the overall general statistics. The estimated results show relative conformity with the LRT annual report (the third-party data) with minor differences, due mostly to the biases of the network data and its features in data collection.

There are several biases and drawbacks when utilizing network data to monitor urban dynamics and travel behaviour. The first one is that the network data is impacted by the usage frequency of cellphone users. Although the communication frequency between network towers and cellphones are more intense in the era of the 4G network than the 3G network or the GSM network, the data quality and update activity interval can still be as low as several hours. In our study, the evening data quantities dropped dramatically with lower usage across all users.

The network system does take some time to update the location information, with the update interval normally around two hours. As a result, the call or text records could be made during a trip but not at the beginning or at the end of the trip. This could potentially mislead the researchers to use the location as a pseudo origin or destination. The temporal resolution has improved with the new 4G network system and does update the location information on a regular basis. Statistically, an average subscriber has 254 records per day, and over 70% of users have one hundred or more records, as shown in FIGURE 4.8.

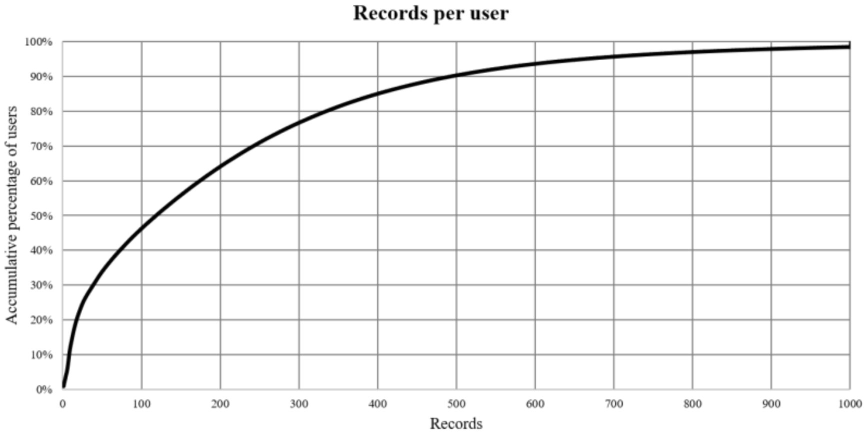


FIGURE 4.8 The cumulative percentage of user's record number

As mentioned above, the network data utilized had a sample rate of 30% of the City of Edmonton's total population. However, not all the samples could be utilized in the estimation since approximately 20% of the population have less than twenty-four daily records. This means that the update interval for these users is greater than one hour. These people are generally considered invalid in the model presented here since the data quality is too low and the update interval is too long to estimate the real travel behaviour, especially as the total travel time from the first LRT station to the last station takes approximately forty-five minutes.

	Identified Total Sample Size	**95% confidence level valid sample size**	**Projected total population based on valid samples**	**Scaling reference from the passenger count number**
Transit Ridership on average	35,789	16,105	120,787	112,805

TABLE 4.2 The identified sample size and 95% confidence level sample size

According to the model and confidence level estimation, 16,105 out of 35,789 records were considered valid samples with a confidence probability higher than 95%. All other records were considered not valid since the model could not make a judgement about the means of transport or whether the person was stationary at points during the entire day. The valid samples were projected to match the total passenger number recorded by the LRT annual report, and all other estimations were based on the projected records.

The following two TABLE 4.3 and 4.4 represent the total origin-destination matrix for all the LRT stations. As shown, the gradient of red represents the intensity of station usage for boarding and alighting during the whole day of operation. The northbound LRT operation matrix shows that most people alighted within the university area during the day, which is intuitively correct since students form a significant percentage of passengers taking the LRT. Furthermore, a significant number of passengers used Grandin station, which linked to key provincial government legislative buildings. The southbound LRT operation matrix shows that the University station was again the most popular boarding station, where students board to go home.

It should be noted that because the LRT track in the downtown area is not covered by the network tower signals, the first or the last on-surface connection is used to determine the station boarded or alighted from by passengers. This estimation could be less accurate than that for the surface stations since the pedways in the downtown area are used by people to travel from one building to another and may distort their true destinations.

As shown in the TABLE 4.2, the surface stations tend to have fewer differences between the estimated passenger boarding and alighting data and the LRT annual report results. These also match with the study's assumption that the data quality is higher for surface stations than for underground stations due to poor connections between network towers and cellphones.

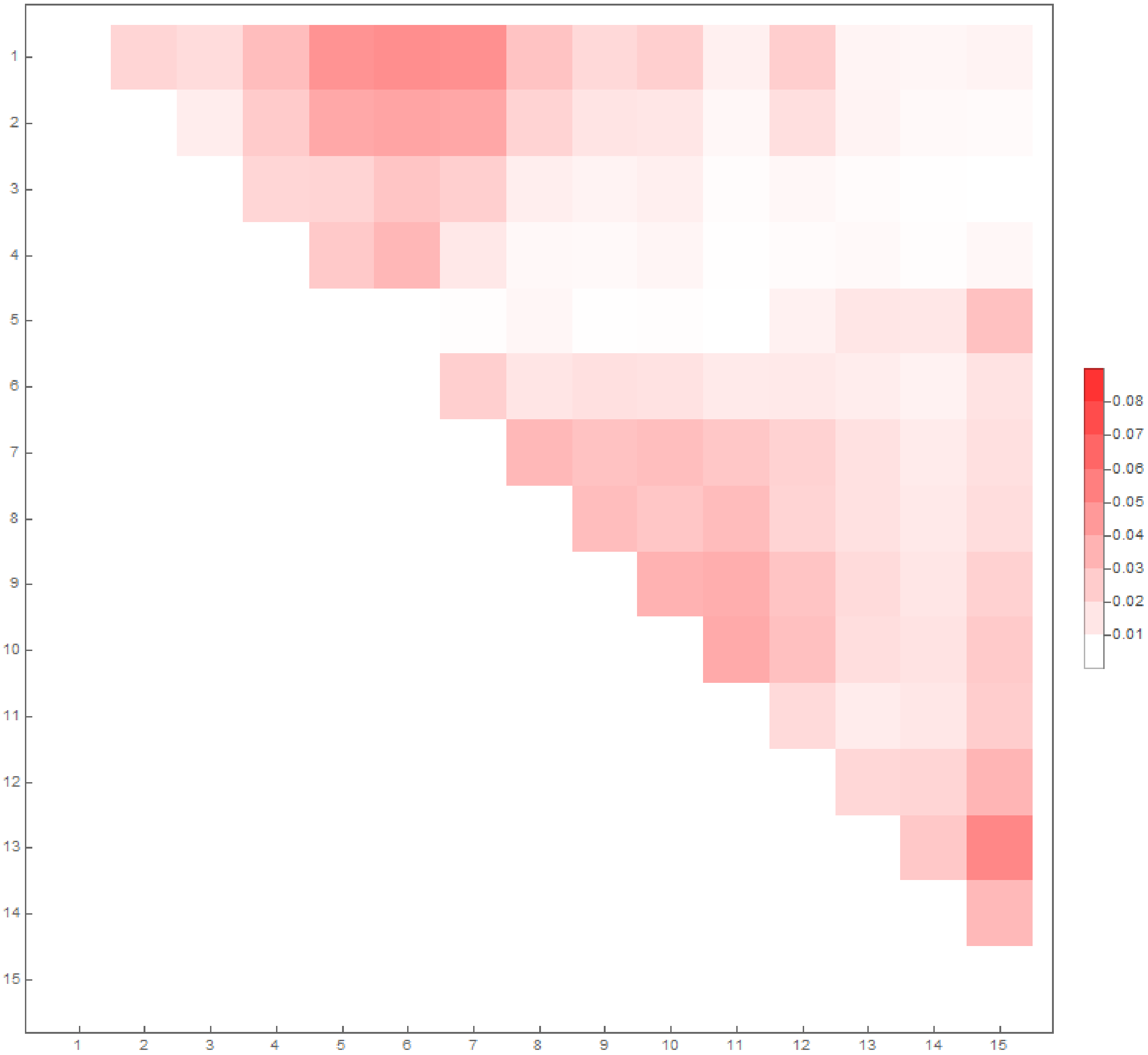


TABLE 4.3 The northbound origin-destination matrix for LRT Capital Line from the valid samples

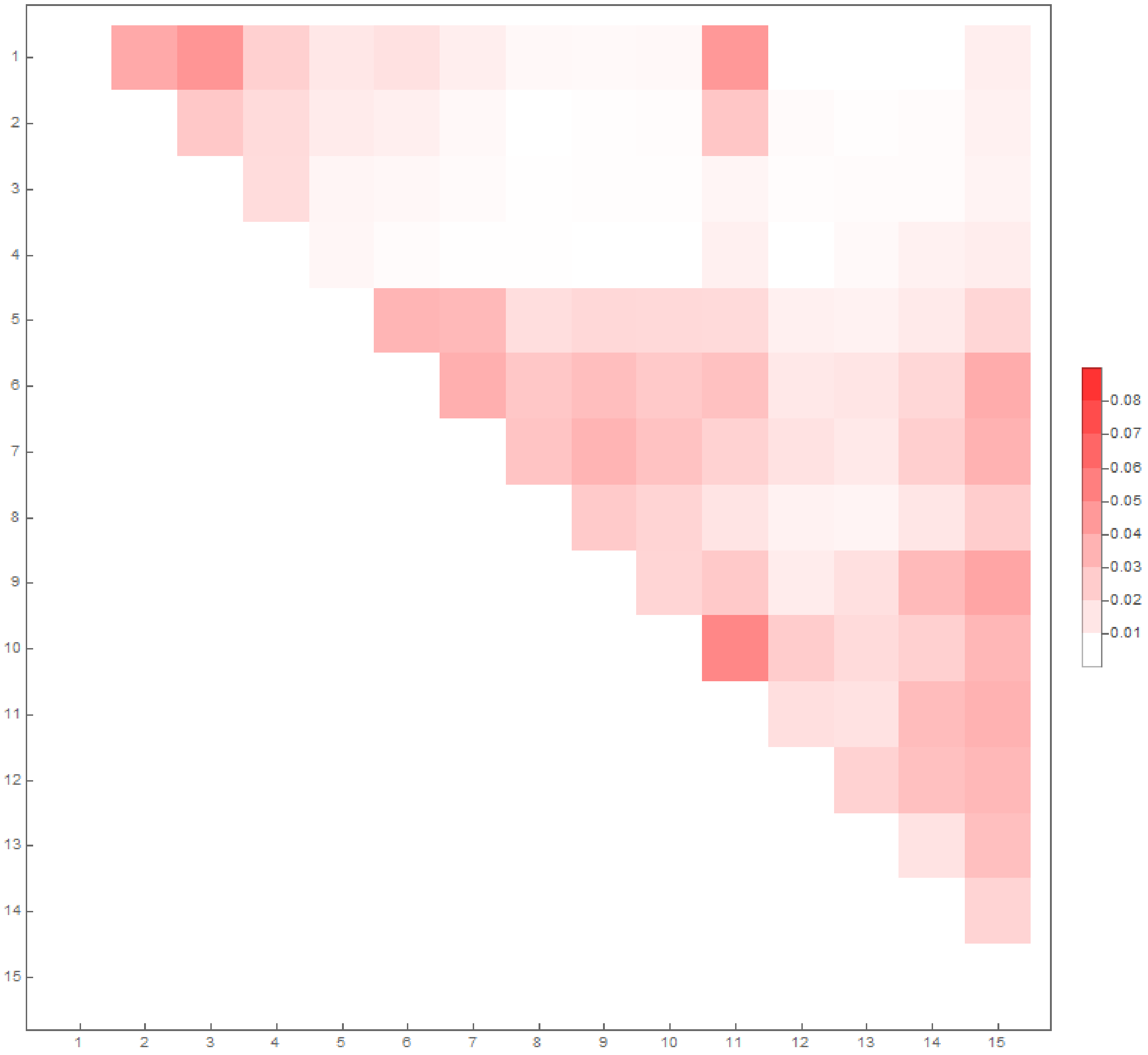


TABLE 4.4 The southbound origin-destination matrix for LRT Capital Line from the valid samples

Number ID	Station Name
1	Century Park
2	Southgate
3	South Campus
4	Belgravia
5	Health Science
6	University
7	Grandin
8	Corona
9	Bay/ Enterprise Square
10	Central
11	Churchill
12	Stadium
13	Coliseum
14	Belvedere
15	Clareview

TABLE 4.5 The station number indication

Both the northbound and southbound origin-destination matrix (TABLE 4.3 and TABLE 4.4) show that both the university and downtown areas are the busiest districts along the LRT Capital line. TABLE 4.5 outlines the Number ID for viewing the origin-destination matrice.

4.7 Hotspot Comparison between Network data and Third-Party Data

The hotspot analysis can help to further understand the validity of the results. In this case, the Century Park station to the University area was chosen for analysis because it is a typical route for students travelling to school and for people using the park and ride to travel to work. The passenger volume changed throughout the day. During the AM peak hours, fewer passengers travelled from the university district to Century Park; during the PM peak hours, large amounts of passengers were captured travelling from university district to Century Park, presumably after work or school.

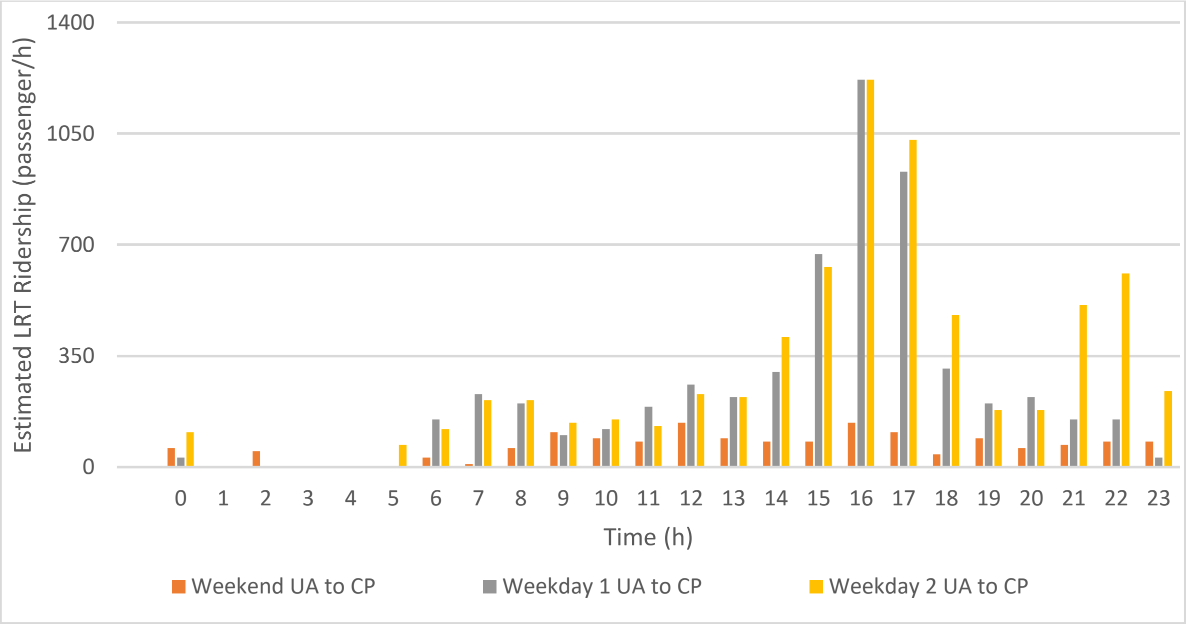


FIGURE 4.9 Hotspot stations trip pattern – the LRT ridership from University to Century Park

CHAPTER 5. CONCLUSION

This chapter gives a summary of the research and its limitations at the current stage. Some thoughts about future work are also discussed in relation to the benefits of this technology on further research and transportation applications.

5.1 Research Summary and Limitations

In this study, a completed data processing architecture was built to capture passenger travel behaviour along the LRT line. A novel method based on pattern matching and trip feature extractions was proposed to infer the LRT trips and non-LRT trips from network data. Network data, transport networks, and travel survey information are jointly implemented in this transport inference model associated with network data travel trajectories. Compared to a traditional travel survey, network data are low-cost, require a less labour-intensive process, and provide up-to-date data information for transportation planning purposes. The proposed model has demonstrated reasonable results in comparison with third-party data and validation methods, and the resulting OD matrices have shown a high correlation with the household travel survey with reasonable differences at certain stations.

Although network data have been more commonly used in identifying the transport mode, the methodology utilized was discussed. The whole data processing architecture included data pre-processing, trip identification, and transport mode identification, where each step required deliberate assumptions to ensure the validity of the output results. As was outlined in the literature review, three types of pre-processing methods have been commonly used in recent studies: pattern-based, time-window-based, and hybrid. The new generation of network data has a higher update frequency and denser cell station distribution, thus increasing the complexity of the raw data. This study has implemented the hybrid method to conduct the data cleaning process, regarded as the most useful method for more complicated trip scenarios.

In terms of trip identification, this study proposed a novel pattern-matching algorithm to estimate the origin-destination flows and transit ridership of the selected LRT line. The pre-defined pattern incorporates the network geolocation, transport networks, and travel survey information that were implemented in the transport mode inference engine. Furthermore, several travel features were extrapolated from each identified candidate trip to estimate the probability of rail travel through a binary logit model. The output results of the OD matrices for the LRT stations on the Capital Line have shown high levels of correlation with reasonable absolute differences, despite some small differences in particular scenarios. Low signal coverage of underground LRT stations in Edmonton meant that the boarding and alighting volumes of the surface stations showed better accuracy and stability than those of underground stations.

5.2 Research Contributions

The output results have demonstrated the capability of capturing everyday travel behaviour. Weekday travel behaviour is highly consistent while it varies with weekend and statutory holiday travel. The travel patterns are also highly repetitive for certain hotspots along the LRT line, such as the University district and Century Park station, where the boarding and alighting volumes are reversed in the morning and afternoon peak hours. In comparison with the field collected GPS data, the model has shown good reliability in identifying long distance rail trips with high confidence.

There are still some limitations to this research. There are no real benchmarks for the OD comparison or for weekday and weekend ridership differences. Therefore, analysis of the results required prior knowledge and common sense. Also, this study focuses on an easy-to-detect transport mode (rail vs. road), which is relatively limited in real world implementation. A more general classification of transport modes would require large numbers of accurate ground truth data, which are difficult to obtain. Considering the advancement of network data technology with improved temporal and spatial resolution, the use benefits of network data to conduct future transport mode studies in large-scale city contexts are increasingly more attractive, particularly in terms of its reduced labour and expense costs. By sufficiently understanding network data, transportation researchers and engineers will be able to improve transport mode identification in large-scale regions and facilitate the investigation of daily city travel.

REFERENCE

1. Gur, Y. J., S. Bekhor, C. Solomon, and L. Kheifits. Intercity Person Trip Tables for Nationwide Transportation Planning in Israel Obtained from Massive Cell Phone Data PLANNING MODEL. Vol. 1996, 1996, pp. 145–151. https://doi.org/10.3141/2121-16.

2. Bachir, D., G. Khodabandelou, V. Gauthier, M. El Yacoubi, and J. Puchinger. Inferring Dynamic Origin-Destination Flows by Transport Mode Using Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 101, No. January, 2019, pp. 254–275. https://doi.org/10.1016/j.trc.2019.02.013.

3. Wang, F., and C. Chen. On Data Processing Required to Derive Mobility Patterns from Passively-Generated Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 87, No. December 2017, 2018, pp. 58–74. https://doi.org/10.1016/j.trc.2017.12.003.

4. Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis. *Transportation Research Part C: Emerging Technologies*, Vol. 68, 2016, pp. 285–299. https://doi.org/10.1016/j.trc.2016.04.005.

5. Chen, C., L. Bian, and J. Ma. From Traces to Trajectories: How Well Can We Guess Activity Locations from Mobile Phone Traces? *Transportation Research Part C: Emerging Technologies*, Vol. 46, 2014, pp. 326–337. https://doi.org/10.1016/j.trc.2014.07.001.

6. Ji, Y., R. G. Mishalani, and M. R. McCord. Transit Passenger Origin-Destination Flow Estimation: Efficiently Combining Onboard Survey and Large Automatic Passenger Count Datasets. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015,

pp. 178–192. https://doi.org/10.1016/j.trc.2015.04.021.

7. Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti. Understanding Individual Mobility Patterns from Urban Sensing Data: A Mobile Phone Trace Example. *Transportation Research Part C: Emerging Technologies*, Vol. 26, 2013, pp. 301–313. https://doi.org/10.1016/j.trc.2012.09.009.

8. Wang, H., F. Calabrese, G. Di Lorenzo, and C. Ratti. Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2010, pp. 318–323. https://doi.org/10.1109/ITSC.2010.5625188.

9. Regt, K. De, O. Cats, N. Van Oort, and H. Van Lint. Investigating Potential Transit Ridership by Fusing Smartcard and Global System for Mobile Communications Data.

10. Eftekhari, H. R., and M. Ghatee. An Inference Engine for Smartphones to Preprocess Data and Detect Stationary and Transportation Modes. *Transportation Research Part C: Emerging Technologies*, Vol. 69, 2016, pp. 313–327. https://doi.org/10.1016/j.trc.2016.06.005.

11. Lan, C. Route-Level Transit Passenger Origin-Destination Trip Estimation from Automatic Passenger Counting Data: A Case Study in Edmonton. 2015.

12. Chen, C., H. Gong, C. Lawson, and E. Bialostozky. Evaluating the Feasibility of a Passive Travel Survey Collection in a Complex Urban Environment: Lessons Learned from the New York City Case Study. *Transportation Research Part A: Policy and Practice*, Vol. 44, No. 10, 2010, pp. 830–840. https://doi.org/10.1016/j.tra.2010.08.004.

13. Luo, D., O. Cats, and H. van Lint. Constructing Transit Origin–Destination Matrices with Spatial Clustering. *Transportation Research Record: Journal of the Transportation*

Research Board, Vol. 2652, 2017, pp. 39–49. https://doi.org/10.3141/2652-05.

14. Rokib S A, Md. Ahsanul Karim, Tony Z. Qiu*, A. K. Origin-Destination Trip Estimation from Anonymous Cell Phone and Foursquare Data. *Transportation Research Record*, Vol. 9, 2015, pp. 1–15.

15. Gundlegård, D., C. Rydergren, N. Breyer, and B. Rajna. Travel Demand Estimation and Network Assignment Based on Cellular Network Data. *Computer Communications*, Vol. 95, 2016, pp. 29–42. https://doi.org/10.1016/j.comcom.2016.04.015.

16. Sadeghvaziri, E., M. B. Rojas, and X. Jin. Exploring the Potential of Mobile Phone Data in Travel Pattern Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2594, 2016, pp. 27–34. https://doi.org/10.3141/2594-04.

17. Zhang, D., J. Huang, Y. Li, F. Zhang, C. Xu, and T. He. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. *Proceedings of the 20th ACM Annual International Conference on Mobile Computing and Networking - MobiCom 2014*, 2014, pp. 201–212. https://doi.org/10.1145/2639108.2639116.

18. Aguiléra, V., S. Allio, V. Benezech, F. Combes, and C. Milion. Using Cell Phone Data to Measure Quality of Service and Passenger Flows of Paris Transit System. *Transportation Research Part C: Emerging Technologies*, Vol. 43, 2014, pp. 198–211. https://doi.org/10.1016/j.trc.2013.11.007.

19. Li, C., C. Zegras, F. Zhao, Z. Qin, A. Shahid, M. Ben-akiva, F. Pereira, and J. Zhao. Enabling Bus Transit Service Quality Co-Monitoring Through Smartphone-Based Platform.

20. Huang, H., Y. Cheng, and R. Weibel. Transport Mode Detection Based on Mobile Phone

Network Data: A Systematic Review. *Transportation Research Part C: Emerging Technologies*, Vol. 101, No. January, 2019, pp. 297–312. https://doi.org/10.1016/j.trc.2019.02.008.

21. Nour, A., B. Hellinga, and J. Casello. Classification of Automobile and Transit Trips from Smartphone Data: Enhancing Accuracy Using Spatial Statistics and GIS. *Journal of Transport Geography*, Vol. 51, 2016, pp. 36–44. https://doi.org/10.1016/j.jtrangeo.2015.11.005.

22. Nour, A., J. Casello, and B. Hellinga. Developing and Optimizing a Transportation Mode Inference Model Utilizing Data from GPS Embedded Smartphones. 2015.

23. Bar-Gera, H. Evaluation of a Cellular Phone-Based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel. *Transportation Research Part C: Emerging Technologies*, Vol. 15, No. 6, 2007, pp. 380–391. https://doi.org/10.1016/j.trc.2007.06.003.

24. Alexander, L., S. Jiang, M. Murga, and M. C. Gonz??lez. Origin-Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 240–250. https://doi.org/10.1016/j.trc.2015.02.018.

25. Iv, M. B. R., E. Sadeghvaziri, and X. Jin. Comprehensive Review of Travel Behavior and Mobility Pattern Studies That Used Mobile Phone Data. No. 2563, 2016, pp. 71–79. https://doi.org/10.3141/2563-11.

26. Kung, K. S., K. Greco, S. Sobolevsky, and C. Ratti. Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data. *PLoS ONE*, Vol. 9, No. 6, 2014. https://doi.org/10.1371/journal.pone.0096180.

27. Shad, S. A., E. Chen, and T. Bao. Resolution Building Cell Oscillation R Esolution in Mobility Profile B Uilding. Vol. 9, No. 3, 2012, pp. 205–213.

28. Calabrese, F., C. Ratti, M. Colonna, P. Lovisolo, and D. Parata. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 12, No. 1, 2011, pp. 141–151.

29. Asakura, Y., and E. Hato. Tracking Survey for Individual Travel Behaviour Using Mobile Communication Instruments. *Transportation Research Part C: Emerging Technologies*, Vol. 12, No. 3-4 SPEC.ISS., 2004, pp. 273–291. https://doi.org/10.1016/j.trc.2004.07.010.

30. Ficek, M., and L. Kencl. Inter-Call Mobility Model: A Spatio-Temporal Refinement of Call Data Records Using a Gaussian Mixture Model. *IEEE Conference on Computer Communications, INFOCOM 2012*, No. Icm, 2012, pp. 469–477. https://doi.org/10.1109/INFCOM.2012.6195786.

31. Zheng, K., Y. Zheng, X. Xie, and X. Zhou. Reducing Uncertainty of Low-Sampling-Rate Trajectories. *Proceedings - International Conference on Data Engineering*, 2012, pp. 1144–1155. https://doi.org/10.1109/ICDE.2012.42.

32. Rose, G. Mobile Phones as Traffic Probes: Practices, Prospects and Issues. *Transport Reviews*, Vol. 26, No. 3, 2006, pp. 275–291. https://doi.org/10.1080/01441640500361108.

33. Calabrese, C., F. Giusy, D. Lorenzo, L. Liu, C. Ratti, F. Calabrese, and G. Di Lorenzo. Estimating Origin-Destination Flows Using Opportunistically Collected Mobile Phone Location Data from One Million Users in Boston Metropolitan Area Terms of Use Estimating Origin-Destination Flows Using Opportunistically Collected Mobile Phone Location Da. *IEEE Pervasive Computing*, Vol. 10, No. 4, 2011, pp. 36–44. https://doi.org/10.1109/mprv.2011.41.

34. Jiang, S., J. Ferreira, and M. C. González. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data : A Case Study of Singapore. *IEEE Transactions on Big Data*, 2015, pp. 1–13. https://doi.org/10.1109/TBDATA.2016.2631141.

35. Bachir, D., G. Khodabandelou, V. Gauthier, M. El Yacoubi, and E. Vachon. Combining Bayesian Inference and Clustering for Transport Mode Detection from Sparse and Noisy Geolocation Data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11053 LNAI, 2019, pp. 569–584. https://doi.org/10.1007/978-3-030-10997-4_35.

36. Çolak, S., L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, and M. C. González. Analyzing Cell Phone Location Data for Urban Travel. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2526, 2015, pp. 126–135. https://doi.org/10.3141/2526-14.

37. Diao, M., Y. Zhu, J. Ferreira, and C. Ratti. Inferring Individual Daily Activities from Mobile Phone Traces: A Boston Example. *Environment and Planning B: Planning and Design*, Vol. 43, No. 5, 2016, pp. 920–940. https://doi.org/10.1177/0265813515600896.

38. Vajakas, T., J. Vajakas, and R. Lillemets. Trajectory Reconstruction from Mobile Positioning Data Using Cell-to-Cell Travel Time Information. *International Journal of Geographical Information Science*, Vol. 29, No. 11, 2015, pp. 1941–1954. https://doi.org/10.1080/13658816.2015.1049540.

39. Bwambale, A., C. F. Choudhury, and S. Hess. Modelling Trip Generation Using Mobile Phone Data: A Latent Demographics Approach. *Journal of Transport Geography*, No. August, 2017, pp. 1–11. https://doi.org/10.1016/j.jtrangeo.2017.08.020.

40. Shad, S. A., and E. Chen. Precise Location Acquisition of Mobility Data Using Using Cell

ID. *Arxiv.Org*, Vol. 1, No. 5, 2013, pp. 5–7.

41. Widhalm, P., Y. Yang, M. Ulm, S. Athavale, and M. C. González. Discovering Urban Activity Patterns in Cell Phone Data. *Transportation*, Vol. 42, No. 4, 2015, pp. 597–623. https://doi.org/10.1007/s11116-015-9598-x.

42. Hoteit, S., S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle. Estimating Human Trajectories and Hotspots through Mobile Phone Data. *Computer Networks*, Vol. 64, 2014, pp. 296–307. https://doi.org/10.1016/j.comnet.2014.02.011.

43. Holleczek, T., L. Yu, J. K. Lee, O. Senn, C. Ratti, and P. Jaillet. Detecting Weak Public Transport Connections from Cellphone and Public Transport Data. *Proceedings of the 2014 International Conference on Big Data Science and Computing - BigDataScience '14*, 2014, pp. 1–2. https://doi.org/10.1145/2640087.2644196.

44. Dong, H., M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, and X. Zhou. Traffic Zone Division Based on Big Data from Mobile Phone Base Stations. *Transportation Research Part C: Emerging Technologies*, Vol. 58, 2015, pp. 278–291. https://doi.org/10.1016/j.trc.2015.06.007.

45. Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. Gonz??lez. Development of Origin-Destination Matrices Using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63–74. https://doi.org/10.1016/j.trc.2014.01.002.

46. Council, D., and N. West. Deriving Operational Origin-Destination Matrices from Large Scale Mobile Phone Data. *International Journal of Transportation Science and Technology*, Vol. 2, 2013, pp. 183–204. https://doi.org/10.1260/2046-0430.2.3.183.

47. Nitsche, P., P. Widhalm, S. Breuss, N. Br??ndle, and P. Maurer. Supporting Large-Scale

Travel Surveys with Smartphones - A Practical Approach. *Transportation Research Part C: Emerging Technologies*, Vol. 43, 2014, pp. 212–221. https://doi.org/10.1016/j.trc.2013.11.005.

48. Järv, O., R. Ahas, and F. Witlox. Understanding Monthly Variability in Human Activity Spaces: A Twelve-Month Study Using Mobile Phone Call Detail Records. *Transportation Research Part C: Emerging Technologies*, Vol. 38, 2014, pp. 122–135. https://doi.org/10.1016/j.trc.2013.11.003.

49. Doyle, J., P. Hung, D. Kelly, and R. Farrell. Utilising Mobile Phone Billing Records for Travel Mode Discovery. *ISSC 2011, Trinity College Dublin, June 23–24*, 2011.

50. Nantes, A., D. Ngoduy, A. Bhaskar, M. Miska, and E. Chung. Real-Time Traffic State Estimation in Urban Corridors from Heterogeneous Data. *Transportation Research Part C: Emerging Technologies*, Vol. 66, 2016, pp. 99–118. https://doi.org/10.1016/j.trc.2015.07.005.

51. Xie, D. X. S. G. C. N. Transportation Modes Identification from Mobile Phone Data Using Probabilistic Models. *International Conference on Advanced Data Mining and Applications*, Vol. ADMA 2011, 2011, pp. 359–371.