Investigating the use of anonymous cellular data for intercity travel patterns in Alberta

by

Tin Ying Hui

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

© Tin Ying Hui, 2017

ABSTRACT

Intercity trips or long-distance trips have been understudied and overlooked by researchers and public agencies in comparison to routine trips that are relatively shorter and often within an urban area. Currently, intercity travel demand is increasing, and accounts for a significant portion of total mileage travelled. This increasing demand also leads to increasing issues such as congestion, energy consumption, and emissions.

Governments are recognizing the need to understand current intercity travel patterns for infrastructure investments and environmental policies, and data such as origin-destination (OD) flows and intercity demand are valuable for strategic planning of the highway networks. However, traditional methods of data collection to estimate OD demands and/or flows through household or roadside surveys are time consuming and expensive, and public agencies have historically prioritized survey data collection within their jurisdiction, typically urban boundaries.

An emerging (albeit imperfect) alternative is passive data sources, such as anonymous cellular data. Anonymous cellular data can provide large random samples with reduced bias and provide results much faster and at much lower cost than travel surveys. It also has a low deployment cost, as it does not require any additional equipment installation or measuring devices. The purpose of this thesis is to investigate how anonymous cellular data may be used to extract more information and features about intercity travel patterns.

In this research, two days of anonymous cellular data for the province of Alberta, Canada are used to extract intercity trips and infer trip modes used. Intercity trips were first extracted between the two major cities - Edmonton and Calgary. The extracted data show that most trips take between 2 – 3 hrs, as well as a smaller portion that take between 0 – 1 hrs. This shows that anonymous cellular data provides a reasonable reflection of the two modes, as a direct drive trip between the two cities take approximately 3 hours, and a flight takes 45 minutes from takeoff to landing. This analysis was expanded to all intercity trips between cities and urban areas in Alberta. Intercity trips between fourteen urban zones in Alberta (including urban service area Fort McMurray and oil sands camps Fort MacKay) were extracted using a similar methodology. Overall, the data shows that larger cities have more trips originate and destined there, and the distance between cities also affected the share of trips (i.e. smaller cities had fewer trips but the highest proportion of trips to and from cities nearby, sparsely located cities had very few trips anywhere).

Two methods were utilized to infer the trip mode for trips between Edmonton Calgary, first by categorizing the travel times using upper and lower limits, second by hierarchical clustering of the travel times. Hierarchical clustering of trips less than 8 hours results in distinct clusters that represent air trips, ground trips, and longer ground trips (likely made with stops). The clusters showed a mode split that ranged from 12 – 25% air trips and 88 – 75% ground trips for the two days. Hierarchical clustering was then conducted for all intercity trip pairs that had direct air service between them, with a unitless, rescaled travel times based on the average ground travel time for each pair. Mode splits ranged widely between the two days of data, which could be due to seasonal variations in trip patterns (i.e. more people flying in winter than in the summer) or sampling issues in the data. The mode split results shows the trend that cities further apart will have a higher share of air trips then ground.

This work contributes to the existing research on intercity travel and passive data applications in transportation. It builds on existing research that have identified origin-destination flows and shows how trip mode can be inferred from travel times, using clustering techniques. This research is limited by the small sample size of two individual days of data, and the data contains only a sample of all records from the cellular service provider. Though the data here is limited, it demonstrates its ability to provide useful information about intercity travel behaviour. Traditional data sources and passive data both have their own limitations, but if used together, they can overcome current limitations in data.

ACKNOWLEDGEMENTS

I would like to give my deepest thanks to my supervisor, Dr. Amy Kim, for without her support my thesis would not be possible. Her guidance, encouragement, and demand for nothing but the best has made me a better student. Also I would like to thank Dr. Tony Qiu, Dr. Karim El-Basyouny, and Dr. Tae Kwon for all their help over the last two years. I would also like to thank my committee members, Dr. Tony Qiu, Dr. Amy Kim, and Dr. Manish Shirgaokar.

To the sponsors and donors that have enabled my studies: NSERC, Government of Alberta, Faculty of Graduate Research Studies, Bruce F Willson, many thanks. As well, special thanks to Edmonton Airports for providing historical flight data used in this thesis.

I would also like to thank all my colleagues that have given support during my time here: Naomi Li, Yunzhuang Zheng, Kexin Ren, Matt Woo, Qianqian Du, Chenhao Wang, Danyang Sun, Kasturi Mahajan, Gloria Duran Castillo, Laura Cabral, Noah Wishart, Suliman Gargoum, Maged Gouda, Mudasser Seraj, Ai Teng, Jiangchen Li, Yuwei Bi, Can Zhang, Chen Qiu, Adekunle Mofolasayo, and Ronald Ekyalimpa.

Special gratitude goes to my parents and my brother, who have always provided me with love and support throughout out my life and during my studies away. And infinite thanks to my best friend, Steve.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Intercity trips and long-distance trips have historically been understudied and largely overlooked by researchers and public agencies. Daily trips that are often relatively short and within urban areas make up the majority of travel compared to more infrequent long-distance trips; due to this collection of long-distance travel data is not prioritized and limited in comparison (Miller, 2004; L. Zhang, Southworth, Xiong, & Sonnenberg, 2012).

Although intercity travel occurs less frequently and regularly than urban commute trips, it accounts for a significant portion of total mileage travelled. For example, in Great Britain, intercity trips make up less than 2% of all trips, but account for 30% of mileage from all trips, urban and long-distance (Dargay & Clark, 2012). Economic growth and increased travel speeds have made it more accessible and faster to take intercity trips; and both intercity trip volume and lengths have increased over the years (Holz-Rau, Scheiner, & Sicks, 2014; Kuhnimhof, Collet, Armoogum, & Madre, 2009). The significant portion of mileage from intercity travel as well as increasing demands leads to larger environmental impacts such as energy consumption, air quality and congestion on intercity corridors (Dargay & Clark, 2012). Governments (such as provinces and states, as well as cities and counties) are recognizing the need to understand current intercity travel patterns for infrastructure investments and environmental policies (L. Zhang et al., 2012), and origin-destination (OD) flows, intercity demands, and traveler characteristics are valuable for strategic planning of intercity transportation facilities (M. H. Wang, Schrock, Broek, & Mulinazzi, 2013). However, traditional methods of data collection for OD flows through household (mail in, phone, internet) or roadside intercept surveys are time consuming and expensive, and public agencies have historically prioritized survey data collection within their jurisdiction, typically urban boundaries (Miller, 2004).

An emerging solution for this lack of intercity OD flow data is to access passive data sources, such as anonymous cellular data. They require minimal equipment and resources, and greater sample sizes can be obtained in less time than household travel surveys. In this thesis I investigate the use of anonymous cellular data to identify intercity travel patterns, specifically OD flows between urban areas in the province of Alberta and their mode split by ground and air. The following section provides more context on intercity travel and its data limitations.

1.1 Thesis Motivation

In the 2001 National Household Travel Survey (NHTS) from the US, they found that 56% of intercity trips are taken for leisure, followed by 16% for business trips and 13% for commuting purposes (Bureau of Transportation Statistics [BTS], 2003). The predominant mode of these trips is by personal vehicle, accounting for 90% of trips, followed by 7% made by air. The NHTS survey estimated that 1.3 trillion person-miles of long-distance travel was made (BTS, 2003), and from the 1995 American Travel Survey long-distance trips accounted for 25% of total personal vehicle mileage (L. Zhang et al., 2012). This large share of intercity trips shows that these trips contribute significantly to congestion, air delays, energy consumption and emissions (Cho, 2013; Kuhnimhof et al., 2009). For example, researchers in Germany found that even though long-distance trips over 100 km accounted for only 2% of all trips, it was responsible for 62% of climate impact from travel (Aamaas, Borken-Kleefeld, & Peters, 2013).

Increasing intercity travel demand generates impacts important to recognize. Public agencies need reliable forecasts to make expensive transportation infrastructure or service decisions (L. Zhang et al., 2012). This demand also leads to higher energy consumption and emissions; a better understanding of intercity travel can assist policy makers to shape environmental policies such as ones that encourage demand shift to higher occupancy modes (which has lower emissions per-passenger-trip) or invest in emission-friendly alternatives (Haobing Liu, Xu, Stockwell, Rodgers, & Guensler, 2016).

However, the largest barrier to this advancement in intercity research is limited data on intercity travel (Anderson & Simkins, 2012). Traditional methods of data collection for OD flows through household surveys or roadside surveys is time consuming and expensive (M. H. Wang et al., 2013) and public agencies have historically prioritized survey data collection within their jurisdiction, typically urban boundaries (Miller, 2004). As people do not make long-distance trips routinely, these surveys would require a larger sample size of respondents or a lengthier questioning period to gather enough long-distance trips for statistical analysis (Kuhnimhof et al., 2009). These surveys are also not able to provide information about the frequency of long-distance trips that users make. As well, long-distance trips are distributed unevenly in the population; some may make long-distance trips often and others rarely (Kuhnimhof et al., 2009).

With this issue in mind, this thesis seeks to investigate how anonymous cellular data, a passive data source, can be applied to identify intercity travel patterns, specifically extracting intercity trips and mode split. In the last few decades, passive data sources such as anonymous cellular data have been used for different transportation applications (Caceres, Wideberg, & Benitez, 2008). Due to the high penetration rate of cell phones – approaching 100% in developed countries – cellular data can provide large random samples with reduced bias and provide results much faster and at much lower cost than travel surveys (Caceres, Wideberg, & Benitez, 2008; Qiu, Jin, Cheng, & Ran, 2007). It also has a low deployment cost, as it does not require any additional equipment installation or measuring devices (Caceres et al., 2008). One of the issues with cellular data involves accuracy, as the exact position of a user's location is not known – only the corresponding network station it is in (Caceres et al., 2008). This becomes a less significant issue in intercity travel, where the regions of interest are larger and less precision is required (M. H. Wang et al., 2013).

Many of the existing studies using passive data focus on the urban scale, covering topics such as identifying traffic state characteristics, travel time, speed, flow, density, OD flows and mode split. Less research has been conducted on intercity travel using passive data; a few studies have estimated intercity OD flows between cities and inferred mode split between different ground transportation modes, which will be discussed in further detail in Section 2.2. My research builds on those studies by exploring the feasibility of utilizing anonymous cellular data to identify intercity travel patterns such as OD flows and mode split.

1.2 Objectives and Tasks

This thesis has two primary objectives in evaluating the feasibility of using anonymous cellular data to gain information about intercity travel patterns. They are:

- To extract intercity trips from an anonymous cellular dataset to identify OD flows between cities in Alberta (not expanded to population volumes), and
- To infer the travel mode (air or ground) of the extracted intercity trips, for travel between cities that are served by air transportation.

Two independent days of anonymous cellular data for the entire province of Alberta are utilized in this study. For the first objective, the data is processed and intercity trips are first extracted between Edmonton and Calgary, then expanded to all cities in the entire province of Alberta. OD matrices of the extracted intercity trips are found, but they are not expanded to population volumes in this thesis. The second objective is conducted by using hierarchical clustering on trip travel times to infer whether a trip is made by air or ground. For intercity trips across the province, the trip mode is inferred if there is air service between the trip's origin and destination. Travel times are rescaled such that all trips with different origins and destinations (and different fly and drive travel times) can be assessed as a single data set.

1.3 Thesis Structure

Chapter 2 is a literature review on the state of the art intercity or long-distance modelling, including common data sources, intercity mode choice models, and intercity demand models. The second half of the literature review discusses the applications of anonymous cellular data in the transportation field, such as travel speed, origin destination (OD) matrices, and mode split. Chapter 3 presents the data processing, methodology, and results of extracting intercity trips. Chapter 4 contains the mode split analysis that is conducted at first by a classification and clustering method between Edmonton and Calgary, then the clustering method is used for intercity trip pairs with air service in Alberta. Lastly, Chapter 5 concludes this thesis report with a summary on the research overview, findings, contribution, limitations and recommendations for future work.

2. LITERATURE REVIEW

The following literature review has three major parts: the first provides background on the current state of intercity travel research with a focus on the data that is commonly used in these studies, while the second reviews the application of anonymous cellular data in urban and intercity transportation applications. Lastly, a brief review is conducted on new alternative sources of data to shed light on other methods and sources that will become more and more relevant in the future. This literature review on intercity travel focuses on the demand side only. Supply issues, such as how a governmental agency allocates funding to intercity highway infrastructure, are out of the scope of this thesis.

2.1 Intercity Travel with Traditional Survey Data

The majority of intercity travel research beginning in the 1950's and 1960's has relied on "traditional" data sources, mainly household travel surveys or choice-based surveys, which ask users about their socio-demographic characteristics and travel, to identify how these characteristics determine travel behaviour (Miller, 2004). Most of these studies revolve around intercity mode choice models, intercity demand models, and specific corridor analyses (to evaluate demand for a new mode such as high speed rail). The state of intercity travel models has not advanced much over the years, with data availability being one of the primary reasons (Anderson & Simkins, 2012; Miller, 2004).

First, the definition of intercity or long-distance travel varies, and is one of the challenges to measure intercity travel (LaMondia, Aultman-Hall, & Greene, 2014). The United States Department of Transportation (USDOT) defined long distance travel trips to be trips 50 miles or longer beginning from home to the farthest destination in the 2001 National Household Travel Survey (NHTS) (USDOT, n.d.). However, in the 1995 American Travel Survey (ATS) it was defined as 100 miles or more. In Canada, the Travel Survey of Residents of Canada (TSRC) considers a long-distance trip to be at least 40 km one-way or an overnight trip (Statistics Canada, 2016). In Europe, a common threshold for a long-distance trip is 100 km (Kuhnimhof et al., 2009). Though these trips are most commonly defined based on a distance threshold, others have defined it based on activity duration, purpose or number of nights away (LaMondia, Aultman-Hall, &

Greene, 2014). From the 2001 NHTS, out of the 2.6 billion trips, 56% of these trips did not include an overnight stay away from home, and 80% of the trips were conducted within the same Census division (Sharp, et al., 2004). From this information, it is apparent that there is no straightforward standard definition of intercity travel around the world.

Despite differences in the definition of intercity travel, the most common form of intercity travel data among all researchers and government officials are household travel surveys. In the United States, prior to 2001 a long-distance travel survey and daily travel survey was conducted separately via the ATS and Nationwide Personal Transportation Survey (NPTS) respectively. The ATS surveyed members in households quarterly asking about long-distance trips they made in the last quarter. Though outdated, it remains to be the primary source of information for travel flows between states and large metropolitan areas (Schiffer, 2012). In the 2001 surveying process, the ATS and NPTS was merged into the NHTS. Respondents were asked to provide information about long-distance trips of 50 miles or more made within the last four weeks. Information about long-distance trips were shortened from a recall period of one year in the ATS to 28 days in the NHTS.

Some states in the USA have also conducted long-distance household travel surveys. Ohio conducted a Statewide Long-Distance Travel Survey in 2002 as a supplement to their Statewide Household Travel Survey in 2001 (Schiffer, 2012). Two-week and four-week retrospective surveys were administered to households to collect long-distance trips that were 50 km or more; the surveys recorded household and person demographics, and travel behaviour characteristics (number of trips, destination, purpose, mode, party size, intermediate stops). Michigan had a long-distance component in their statewide travel surveys (in 2004 and 2009), which defined long-distance trips as 100 miles or greater and a recall period of 3 months before the survey (Schiffer, 2012). The 2010–2012 California Household Travel Survey also contained a long distance component, where respondents were asked to log any trips 50 miles or more made within eight weeks of the survey (Kunzmann & Daigler, 2013). Similarly, other countries also use some variant of a national travel survey to collect information regarding intercity travel. For example, in Great Britain, intercity travel was found to account for 30% of the total distance travelled made from all urban and long-distance trips in the country despite that only 2% of trips were long-distance trips (Dargay & Clark, 2012). In this study, researchers also found that long-distance travel is most elastic to income, where people with higher income will make more long-distance trips.

In Canada, travel surveys measuring leisure travel is broken into domestic and international travel. Domestic travel is measured annually by the Travel Survey of Residents of Canada (TSRC) since 2005, replacing the Canadian Travel Survey (Statistics Canada, 2017c). The purpose of the TSRC is to capture domestic tourism; which includes leisure trips, business trips that are not part of their daily routines, and other miscellaneous trips. The TSRC captures out-of-town overnight trips, as well as out-of-town same day trips of at least 40 km from home, with a recall period of one month (Statistics Canada, 2007).

Other than home-based sampling (such as household travel surveys), the other common form of traveler survey is choice-based sampling (Miller, 2004). Choice-based sampling targets respondents that have already chosen to use a certain mode, common for analyzing a specific travel corridor, with interest to invest in a new travel mode (Miller, 2004). These studies primarily focus on developing mode split models to provide an estimate of new mode shares. For example, in Alberta, a feasibility assessment to consider high speed rail (HSR) in the Edmonton-Calgary corridor conducted surveys at the Edmonton and Calgary airports, bus terminals, and identified auto drivers by recording their license plates on video camera and mailing surveys out to them (TEMS, Inc. and Oliver Wyman, 2008). The consultants mailed out a stated preference (SP) survey, which asks about hypothetical choices, such as whether a respondent will choose to take bus, rail, auto, or HSR if it exists, and under sets of specifically designed conditions (Bradley & Daly, 1997). Surveys that reveal the choice that respondents actually make are known as revealed preference (RP) surveys.

A study conducted by LaMondia, Moore, and Aultman-Hall (2015) used a one year panel set of overnight trips to model the time interval between individuals making overnight long-distance trips. This is one of the few studies that have looked at patterns or frequency of people making long-distance trips; this data is not typically collected in household travel surveys due to their short recall periods.

2.1.1 *Disaggregate Intercity Mode Choice Models*

The primary focus of intercity research has been on mode split analysis, often because there is interest to introduce a new mode on a travel corridor (Miller, 2004). This research takes in the form of discrete choice models, where a choice is made from a finite set of alternatives (Anderson & Simkins, 2012; Train, 2009).

On a national level in the United States, four major attempts have been made to develop an intercity mode choice model between 1976 and 1990 (Ashiabor, Baik, & Trani, 2007). The first was completed in 1976 in the form of a multinomial logit model using the 1972 National Travel Survey (NTS) (Stopher & Prashker, 1976). The second was in 1981, when a multinomial logit model was constructed using the 1977 NTS data (Grayson, 1981). The third by Morrison and Winston in 1985 with a nested logit model using the same 1977 NTS database (Morrison & Winston, 1985), and later in 1990 Koppelman extended Morrison and Winston's work (Koppelman, 1989). In all four studies, the intercity trips were restricted to intercity trips made between metropolitan statistical areas (MSAs) as the survey only collects the state and MSA (if applicable) of a trip's origin and destination.

Most recently, Ashiabor et al. (2007) utilized the 1995 ATS to create a mode choice model to estimate market share between automobile and airports between any two counties or airports in the United States. This was the first attempt at measuring county-to-county demand versus demand between urban areas with satisfactory results.

Various disaggregate mode choice models have also been applied to survey data collected in Canada. The earliest one was conducted by Transport Canada in 1976 and another by Ridout and Miller in 1989 (Abdelwahab, 1991). Ridout and Miller explored disaggregate logit models before they were widely used, using the 1969 Canadian Transport Commission (CTC) survey data (Ridout & Miller, 1989). This was a choice-based survey that surveyed people travelling along the Windsor – Quebec City corridor on three different modes: air, bus, and rail. The survey asked about their personal demographics and trip details. The respondent's origins and destinations were coded to their census tracts, which was more detailed than the 1980 Canadian Travel Survey (CTS), which coded origins to a CMA and destinations to a census subdivision or division.

In 1991, Abdelwahab used the CTS of 1984 to create two intercity mode choice models for Canada; one for the eastern region and another for the western region (Abdelwahab, 1991). Only trips made from CMA to CMA could be analyzed (which was 9% of all the trips recorded in the survey), as the census divisions were too large to pinpoint an origin and destination. He tested the transferability of one model to another with poor results.

A multinomial logit model of intercity mode choice was developed using data from the CTS. All intercity trips made between 23 CMAs in Canada were extracted to be used in the model (Wilson, Damodaran, & Innes, 1990). They identified many limitations to the CTS data that makes it difficult for intercity travel analysis. Like other studies, they found that the geographic detail of the data was too macroscopic; any intercity trips made in rural areas were categorized into their census division. As well, the data did not provide household car ownership and the number of household members owning a drivers' license.

In 1995, Bhat created a heteroscedastic extreme value model of an intercity travel mode choice using a 1989 Rail Passenger Review survey conducted by VIA Rail. This survey was conducted to develop intercity travel demand models and estimate the mode shift effects of improving potential rail services on the Toronto – Montreal corridor. Among multinomial and nested logit models, he found that the heteroscedastic extreme value model performed the best and had the advantage of overcoming the independence of irrelevant alternatives property over the multinomial logit model (Bhat, 1995).

Koppelman and Wen used the same data set from VIA Rail to develop other mode choice model structures. From this, they developed a paired combinatorial logit model as well as a generalized nested logit model (Koppelman & Wen, 2000; Wen & Koppelman, 2001).

LaMondia, Aultman-Hall, & Greene (2014) conducted a retrospective online survey, which asked participants on their long-distance travel behaviour for the past year. The difference between the long-distance trips asked here versus the NHTS is that there was no distance threshold for a trip to be considered long distance. Using an ordered probit regression model, they found that education level and income were most correlated with an increase to both work and non-work travel, while having a spouse and children decreased some types of long distance travel.

2.1.2 *Intercity Travel Demand Models*

Total demand models, or direct demand models, estimate the volume of travel between an origin and destination. The majority of intercity models in the field currently use these models to estimate demands; however they are outdated with several issues in the functional form (Miller, 2004). They are commonly expressed in the form of a Cobb-Douglas function, where socio-economic factors in the origin and destination zone affect the demand between them (Miller, 2004). They develop aggregate passenger demands between origins and destinations, which give them the advantage of being less data intensive compared with disaggregate models (L. Zhang et al., 2012).

Oum and Gillen (1983) investigated the structure of intercity travel demands in Canada. They developed direct demand functions for three different passenger modes: bus, rail, and air. An HSR assessment conducted in the Edmonton-Calgary corridor utilized a total demand model to estimate demand between the corridor; for trip purposes: business and other, using socioeconomic characteristics: population, employment, and income (TEMS, Inc. and Oliver Wyman, 2008).

2.2 Anonymous Cellular Data in Transportation

There has been much research conducted in recent years on transportation applications using anonymous cellular data, with the bulk focused on understanding traffic states, applications within the four-step model (ex. origin and destination matrices, trip distribution), mode splits, and human behaviour.

Anonymous cellular data exists because of communications between the cell phone and the cellular network. The cellular network is called the Global System for Mobile Communications (GSM) (Schlaich, Otterstatter, & Friedrich, 2010). A typical GSM cellular network consists of base transceiver stations (BTSs), which are where cell phones send and receive their signals. These BTS's are grouped and controlled by Mobile Switching Centers (MSCs). The coverage of a BTS is usually divided into several areas called cells, and the area covered by a MSC is called a Location Area (LA). Each cell and LA has a unique identification (ID). An MSC contains a database that records and stores a mobile users' activities or signaling events in the cellular network. The different types of signaling events are listed below:

- Mobile originate call/SMS: make a phone call or send a message
- Mobile terminated call/SMS: answer a phone call or receive a message
- Handover: switch to another cell during a call
- International Mobile Subscriber Identity (IMSI) attach: cellphone switched on
- IMSI detach: cellphone switched off
- Location update: switch to another LA
- Periodic location update: idle for a period of time, default is often every 2 hrs

In the raw form, anonymous cellular data will contain a list of records, with the following information: anonymous user ID, timestamp, cell ID, LA ID, and event. This data is rich as it contains all communications between cell towers and cell phones (Caceres, Romero, & Benitez, 2012; Schlaich et al., 2010). On the other hand, a more processed form of anonymous cellular data is also available known as Call Detail Records (CDR), created for billing purposes (Horak, 2006; Steenbruggen, Tranos, & Nijkamp, 2015). Each CDR for a user contains entries for billable cellular activity such as calls, text messages, and data usage, with details of the call/text duration, start time, and location, identified by a cell tower ID (Horak, 2006; Steenbruggen et al., 2015).

2.2.1 ***Travel Speeds***

A review of cellular data in transportation applications by Caceres et al. (2008) identified that anonymous cellular data has most commonly been used to identify traffic speeds, traffic volume, travel time, congestion, and density, which are all traffic characteristics.

Studies that utilize anonymous cellular data to measure travel time or speed measurements began in the mid 1990's (Qiu et al., 2007). One study in Lyons, France observed travel speeds using anonymous cellular data on the Rhone Corridor as part of a field test (Ygnace, 2001). The authors compared travel times from anonymous cellular data with loop detector data and found that loop detectors recorded higher speeds.

Bar Gera (2007) analyzed the feasibility of using anonymous cellular data to obtain travel time estimates. He focused on anonymous cellular data, by analyzing the sequence of the handover events. Travel time estimates are compared with loop detector data, for a segment of a busy highway in Israel. The author concluded that the travel time estimates from anonymous cellular data provided good results; however the results showed that the anonymous cellular data has much more noise than the loop detector data.

An American startup company called AirSage, Inc. has developed a system to analyze travel times on freeways and arterials in Minneapolis, Minnesota (Henry X. Liu, Danczyk, Brewer, & Starr, 2008) by using Spring PCS's cellular network. The Minnesota Department of Transportation consulted researchers to assess the accuracy of AirSage's travel times and speed estimates, by comparing their results with field data. They found that results varied in different conditions, and would depend on what are acceptable margins of errors.

Other researchers further studied using handover events to measure traffic speeds; they found that cellular phone probes perform better in free-flow conditions than in congested areas, and intersection delays on local roads can make speed estimates much less accurate (Fei Yang, Yao, & Yang, 2016).

2.2.2 *Origin-Destination (OD) Flows*

It is crucial to understand how and where people move or travel (demand) so that the transportation infrastructure (supply) is able to handle these volumes efficiently. There have been studies that utilized anonymous cellular data to study human trajectories (Gonzalez, Hidalgo, & Barabasi, 2008; Hoteit, Secci, Sobolevsky, Ratti, & Pujolle, 2014). Anonymous cellular data has also been applied to estimate the volume of trips between origins and destinations mostly within cities. These OD flows are important for strategic planning of transportation networks and to identify where infrastructure improvements are required (Calabrese, Di Lorenzo, Liu, & Ratti, 2011). They estimate current demands and can be input as part of the four-step model to forecast future demands (Miller, 2004). OD flows or matrices have been estimated with limitations via surveys or traffic counts (Y. Zhang, Qin, Dong, & Ran, 2010). Studies within the last decade have instead began using anonymous cellular data to estimate these OD matrices which are explained below.

Earlier studies used simulated cellular data to study OD flows (Caceres et al., 2007). Travelers (i.e. vehicles) are simulated using traffic microsimulation programs such as VISSUM, with a module that can simulate cell phone signaling events (K. Sohn & Kim, 2008; Y. Zhang et al., 2010).

In 2007, Caceres et al. (2007) utilized simulated cellular data of location updates to obtain OD flows along a highway corridor. The highway corridor between two Spanish cities was divided into four zones as characterized by the LA of the cellular network. They simulated vehicles (containing mobile users) travelling along the corridor for one day and obtained OD flows by taking the first and last LA in which there was a location update as the origin and destination.

K. Sohn and Kim (2008) estimated OD flows indirectly with handover events. They used VISSUM to simulate traffic count and cellular data for a network of interchanges in the northeastern part of Seoul, South Korea. The handover event times that occurred entering and leaving a cell were used to estimate an approximate time the user was in the cell, to obtain a traffic "count" from the cellular data.

Y. Zhang et al. (2010) later used a methodology that included all signal events from simulated cellular data (including location updates and handover events) in estimating an OD matrix. They simulated traffic and cellular data in Madison, Wisconsin using household data, then estimated an OD matrix using their methodology to compare the results from the OD matrix from household survey data.

The first study that used actual anonymous cellular data for OD flows was conducted by researchers that used a data set of over 1 million users in the Boston metropolitan area provided by AirSage Inc. (Calabrese et al., 2011). AirSage, Inc. provided a triangulated estimate of the location within a cell. The authors concluded that this data provided valuable OD estimates and are advantageous in that weekday and weekend OD flows could be found.

In Madrid, researchers inferred OD matrices by using anonymous cellular data to update a prior matrix, which is a previously estimated matrix inferred from survey data. In order for traffic analysis zones (TAZs) to match up with cell boundaries, links with boundary crossings were grouped together to create aggregate volumes (Caceres, Romero, & Benitez, 2013).

Iqbal, Choudhury, Wang, & Gonzalez (2014) utilized over a month of CDRs to estimate an OD matrix of zones within Dhaka, Bangladesh. Their methodology combined anonymous cell phone data with limited video vehicle counts to create an OD matrix, by using a simulator to create additional traffic counts based on the video vehicle counts. The OD matrix from the CDR data was expanded using a scaling factor from the simulated traffic counts.

Colak, Alexander, Alyim, Mehndiratta, & Gonzalez (2015) used several months of CDR data from Boston and Rio de Janeiro to develop a methodology to estimate urban OD matrices by first detecting stays and then assigning them as home, work or other. Alexander, Jiang, Murga, & Gonzalez (2015) also developed a methodology for creating OD matrices by looking at the frequency of visits in certain time intervals to assign trip purpose, with the results scaled up and compared with census and travel survey data.

There have been studies that have identified OD flows between different cities. In Germany, a study utilized anonymous cellular data with traffic counts to create OD matrices across a network of long-distance highway interchanges (Schlaich et al., 2010). In 2013, a pilot study in Israel created an OD matrix between urban areas on a national scale using records from over 900 cellular phone numbers for a week (Bekhor, Cohen, & Solomon, 2013).

M. H. Wang et al. (2013) estimated OD traffic flows and demand between three cities in the Kansas Metro Corridor using anonymous cellular data provided from AirSage Inc. Six weeks of data was provided, however the records were only generated when cell phones made a phone call, texted, or connected to the Internet.

2.2.3 *Mode Split*

A small number of studies have also utilized anonymous cellular data to study mode split. In 2006, T. Sohn et al. (2006) utilized anonymous cellular data to infer a person's mobility state as one of: stationary, walking or driving. A real-time study analyzed the movement of people in Rome; an algorithm was used to infer whether a cell phone user was walking or in a vehicle during a call (Calabrese & Ratti, Real Time Rome, 2006). A study in Germany generated trajectories using location area updates in southwest Germany for a highway network (Schlaich et al., 2010). They included an analysis that differentiated between private automobiles, trucks (slower vehicles), and trips made with stops using an algorithm, though the algorithm is not explained. H. Wang, Calabrese, Di Lorenzo, & Ratti (2010) inferred trip mode with one month of CDR data provided by AirSage in Middlesex County, Massachusetts. They identified users driving, taking public transit, or walking for different origins and destinations by applying a k-means clustering technique on the travel times (H. Wang et al., 2010). This is the only research thus far to have utilized only the travel times to infer trip mode. Trips were grouped into origins and destinations of 500 m x 500 m wide cells. They defined the errors of the transportation mode inference to be the average distance between the average travel times from each cluster and the corresponding Google Maps travel time by mode. They evaluated the clustering performance by looking at the error, finding on average an error of 5-6 minutes, as well as using the silhouette value, a common method to identify whether data in each cluster are well associated with each other, which performed fairly well.

2.3 Additional Sources of Data

In addition to traditional sources of data and data collection such as household travel surveys, and much explored passive data sources such as GPS, Bluetooth, and anonymous cellular data, now there are an increasing number of opportunities to collect data that can tell us something about transportation patterns and behaviours. These opportunities may arise due to technological advances as well as merging different and new data sources (such as passive data sources, the

Internet, social media) with traditional methods of data collection. One example of this is by using smartphones to collect trips passively that users make, which the smartphone can later ask users on the trip details and collect the user's demographic information. For example, the Ohio Moves Transportation Study used a smartphone app for a few months that detected when they made a trip longer than 50 miles, then the app would ask the user a few questions on those long-distance trips (Ritter & Greene, 2017). Another advantage of this method is that the app can measure the frequency of long-distance trips made, which is often a limitation to long-distance trips collected in surveys (studied by LaMondia, Moore, & Aultman-Hall (2015)). Similarly, a study in Japan and another in Sydney utilized GPS equipped cell phones to record trips, and users would later fill out an internet travel diary that asked targeted questions about the trips they made (Itsubo & Hato, 2006; Stopher & Collins, 2005).

Other than our phones, various other technological advances will be able to provide innovative advances in collecting travel data. For example, Internet giant Baidu is tracking the location data of their users to identify areas of highest internet traffic in China (Hodson, 2016). Uber has begun to release data that shows how long it takes to get from one point to another depending on the time of day and day of week (Poon, 2017). Social media platforms such as Twitter and Foursquare are being mined by researchers to study the trajectories of humans geographically in daily life (Badger, 2013; Jurdak, et al., 2015), during disasters (Sakaki, Okazaki, & Matsuo, 2010; R. Q. Wang & Taylor, 2014), and more. A handful of studies have used Foursquare data, a location-based social network where users can let their network know where they have been by checking in. Some studies have focused on using Foursquare to estimate OD flows or demand (S A, Karim, Qiu, & Amy, 2015; Fan Yang, Jin, Wan, & Li, 2013). For example, a study in Edmonton uses Foursquare data and anonymous cellular data to estimate OD flows and compared to an existing OD matrix from travel surveys (S A, Karim, Qiu, & Amy, 2015). Another study combines anonymous cellular data with Foursquare data, and with machine learning, predicts the type of activity of a Foursquare user's check-in at certain venues (Noulas, Mascolo, & Frias-Martinez, 2013).

Apart from common passive data sources like GPS or anonymous cellular data, and traditional household or choice-based surveys, there have been an increasing number of methods to collect meaningful transportation data. In the context of this thesis, it is important to be aware of this as many of the limitations of the anonymous cellular data in this thesis could be supplemented with other methods of collection in the future to provide a more robust dataset.

2.4 Conclusions

This literature review focused on three major areas of research: the state of existing intercity travel research, anonymous cellular data in transportation applications and new alternative sources of data. I find that there is no standard definition of long-distance or intercity travel; many long-distance surveys use a distance threshold to categorize long-distance trips, but they can also be defined by duration or purpose. So far the primary source of long-distance travel are national level surveys such as the 1995 American Travel Survey (which no longer exists), the National Household Travel Survey, and the Travel Survey of Residents of Canada. Researchers have found that these surveys are limited in many ways, for example, the smallest unit of aggregation is at the CMA level, and it is impossible to study any trips that originate or end outside of a CMA (census subdivision or division level is too aggregate). I find that anonymous cellular data has been used for many applications, of particular interest OD flows and mode split. The quality and source of data varied greatly in these studies, including: simulated data, CDRs, data from a second hand party such as AirSage, Inc., or anonymous cellular data directly from the service provider.

Survey data and passive data both have their advantages and limitations. Household travel surveys have served as the primary data source that provides detailed information about the trip maker (Dargay & Clark, 2012). A well recognized limitation to passive data sources is that there is no knowledge about the trip maker or trip purpose, and many agree that passive data cannot replace “traditional” methods of surveying (Lee, Sener, & Mullins III, 2016). Survey data is needed to gather a user’s socioeconomic demographic characteristics so that travel models can associate trip behaviours with these characteristics (serving as explanatory variables of travel choice). Passive data has the advantage that typically, much larger sample sizes can be obtained without much additional cost; if passive data can provide information about a user’s home location (such as TAZ or electoral ward), then aggregate socio-demographics based on the home location may be linked to (groups of) users. Although this thesis focuses on passive data specifically anonymous cellular data, it is important to be aware of the innovative potential methods of combining passive data sources with data collected in “traditional” survey methods.

This research builds upon existing research in multiple ways. First, the majority of work studying OD flows have focused on intra-urban travel and others that involve intercity OD flows have done so with cities closely connected by highways. I identify OD flows between fourteen urban zones in the province of Alberta, which is a much larger geographic scope than what other studies have covered thus far with anonymous cellular data. Only two of these fourteen urban zones are CMAs; my work can help show how anonymous cellular data can be used to study intercity travel on a more detailed geographic level that surveys cannot. As well, I build upon mode split work that H. Wang et al. (2010) has done, by clustering with travel times, but on an intercity level with modes air and ground. A novel part of this analysis includes clustering trips from all OD pairs together by normalizing the trip travel time.

3. Extracting Intercity Trips from Anonymous Cellular Data

In this chapter, two days of anonymous cellular data are utilized to extract intercity trips in the province of Alberta, Canada, beginning with an analysis between the two primary cities Edmonton and Calgary, then further expanded to cities in the province. The extracted trips simply represent trips in the data and are not expanded population volumes, which is out of the scope of this thesis. The location background, data processing, and trip extraction methodology are provided in detail in this chapter.

3.1 Extracting Intercity Trips between Edmonton and Calgary

Edmonton and Calgary has one primary corridor connecting the two cities and direct air service between them. The primary driving route and direct air service make it a suitable city pair to extract trips and conduct a mode split analysis (conducted later in Chapter 4). In the analysis, an initial focus was first put on Edmonton and Calgary as they are expected to be responsible for a significant portion of travel in Alberta, due to their high populations.

3.1.1 *Background*

As of 2017, Alberta had a population of 4.29 million people (Alberta Government, 2017). In 2011, 83% of the population lived in urban areas and 17% in rural areas (Statistics Canada, 2011). The City of Edmonton and the City of Calgary are the two largest cities in Alberta, with Edmonton being the provincial capital. They are also the only two cities with a Census Metropolitan Area (CMA) classification by Statistics Canada. Their city populations (not CMA) are: Edmonton: 899,400 (Election and Census Services, 2017), and Calgary: 1,239,220 (The Canadian Press, 2017).

The two cities are connected by Highway 2 (Hwy 2), a provincial highway and core route in the National Highway System in Canada (Figure 2) (Stantec, 2007). The cities are also connected by direct air services; Edmonton International Airport (YEG) had 7,466,141 total passengers enplaned and deplaned in 2015 and Calgary International Airport (YYC) had 14,578,929 (Statistics Canada, n.d.). Despite the economic downturn in 2014, YYC has continued to grow in annual passengers

(CBC News, 2016) and YEG has declined (and growing slowly again) (Edmonton International Airport, 2016). It should be noted here that YYC is a hub airport while YEG is not a hub but, rather a "focus" airport for the two largest Canadian air carriers (Westjet and Air Canada); YYC has more service and at higher flight frequencies than YEG.

The approximate travel time by private automobile between Edmonton and Calgary ranges from 2.75 to 3.5 hrs according to Google Maps depending on the start and end location within each city. Express busses take four hours with non-express service taking over five hours. The average flight time from take-off to landing between YEG and YYC averages 43.1 minutes as obtained from historical flight data (explained later in Section 4.1.3).

3.1.2 *Data Processing and Trip Extraction Methodology*

Samples of anonymous cellular data were provided for one day in summer (Day 1) and another day in winter (Day 2) across the Province of Alberta, within the last five years. The raw data contains approximately records for 300,000 users on each day. Each record includes the encrypted cellphone ID (namely the IMSI), timestamp, cell ID, LA ID and signaling event ID. There is no continuity between user IDs from Day 1 to Day 2. Alberta has the highest mobile telephone penetration rate in Canada at 90.1%, compared to the entire nation at 81.4% in 2012 (Canadian Radio-television and Telecommunications Commission, 2014). The raw anonymous cellular data was verified between Edmonton and Calgary using field tests, where the raw data for a specific cell phone was verified with the actual locations the cell phone user travelled to.

Database software was used to preprocess the raw cell phone data by removing erroneous and irrelevant entries. First, any entries for an anonymous user that were missing cell ID's or LA ID's were removed. Secondly, the database software also removed any excess entries showing the "flip flop" effect, which occurs when a stationary phone will flip between cell towers frequently because the phone is on the border of different cell towers.

Prior to extracting intercity trips, an analysis was conducted to identify whether the data is sampled consistently and representative of the population. All users that had records in Edmonton were identified. Users' cell ID or LA ID during typical at-home hours were aggregated to the city's 31 large level Traffic Analysis Zones (TAZs). This analysis compares the population distribution (from 2001 Census of Canada) of people's homes by TAZ with the sample distribution of user's homes from the data. For each applicable user, the home TAZ was determined to be the zone with the longest stay time in total from the periods: 12 – 8 am, and 6 – 11:59 pm. The comparison is shown in Figure 1.

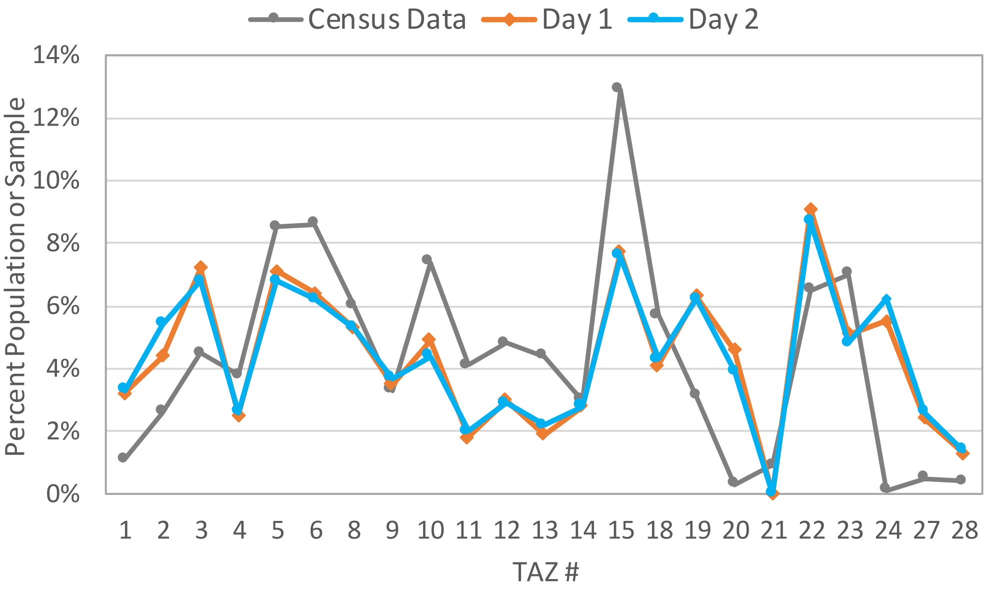


Figure 1 Comparison of population distribution from census data and anonymous cellular data. Census data from: (City of Edmonton, 2017).

As seen in Figure 1, the sample distributions for both days are quite similar to each other, although there is more variation between the census data population distributions and the two days of anonymous cellular data. This suggests that the data between the two days are sampled consistently geographically. The census data is from 2001, which is over ten years older than the samples from the anonymous cellular data, which may account for some of the inconsistency. Other differences may be due to sampling bias of anonymous cellular data; TAZs with certain sociodemographic characteristics may be over or under-represented. However, the overall patterns of distribution from the data may be considered reasonably similar to the census data, and further data processing is conducted to begin extracting intercity trips, as described below.

The spatial resolution of the anonymous cellular data was aggregated from the cells or LAs to large, city-scale zones to identify intercity trips. These zones (Figure 2) were: Edmonton CMA, City of Edmonton, Calgary CMA, City of Calgary, Hwy 2 north of Red Deer, Hwy 2 south of Red Deer, and Red Deer.

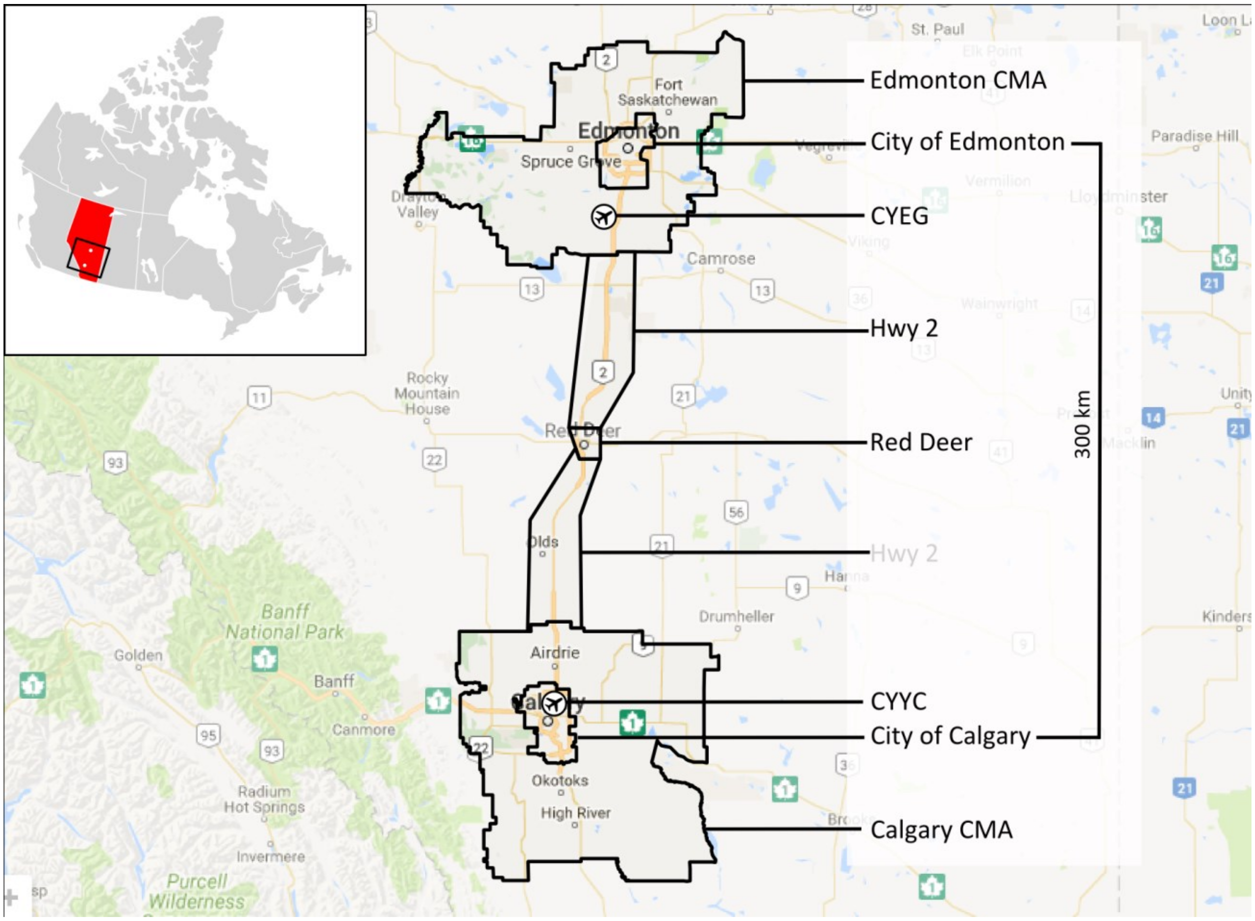


Figure 2 Region of interest in Alberta Canada with two largest cities.

The database filters and identifies all users that had at least one entry within Calgary CMA (city included) and one entry in the Edmonton CMA (city included). Thus, only the records for users that made an intercity trip between Edmonton and Calgary are identified.

The data was processed to output only the first and last entry in each large zone ID. The first and last entries represent minimum boundary times that the user was within the respective zone. For example, a user sends five text messages within the City of Edmonton and then begins driving south to Calgary. The processed data could possibly have the following entries: the first showing location of City of Edmonton, first timestamp being the time the first text message was sent, and last timestamp being the fifth text message sent. The next entry for this user could possibly be a location update on Highway 2 or Red Deer. Later there would be an entry showing City of Calgary, the first observed cellular activity in Calgary, and the last observed activity there.

This processed and cleaned data – consisting of Anonymous User ID, Location Zone, First Timestamp, and Last Timestamp – was output to comma separated delimited (csv) files for further analysis. On Day 1 there were 8282 records with 1307 unique ID's, and 8332 records with 1635 unique ID's on Day 2.

A preliminary look at the data identified that a small portion of users' records contained entries in the following pattern: city 1-city 2-city 1 in an impossible timeframe, most often with less than fifteen minutes between entries. These were identified as data errors which required removal. If a user traveled to city 2 in less than half an hour and stayed in city 2 for less than one hour, then this trajectory was deemed to be invalid and removed. On Day 1, 5% of users were removed and 20% of users on Day 2.

Once all users with entries in Calgary and Edmonton were identified, and any known errors filtered out, the trip extraction was performed as follows:

- For a user, the first record (chronologically) in Edmonton or Calgary was assigned as the Start Location, where the Last Timestamp in that zone assigned as the trip Start Time. For Edmonton, this included the large zones: City of Edmonton, Edmonton CMA, and YEG, and for Calgary: City of Calgary, Calgary CMA, and YYC.
- The other city that was not the Start Location city is then assigned as the End Location. The First Timestamp that occurs in the End Location is assigned as the trip End Time.
- A binary variable called Between was introduced, with a value of 1 if there were any records on Hwy 2 or Red Deer the trip in between Start Time and End Time
- The trip Travel Time was calculated as the time difference between the End Time and Start Time.
- Record Anonymous User ID, Trip Start Location, Trip End Location, Trip Start Time, Trip End Time, Trip Travel Time, Between in a Matlab structure array.
- Loop through every user to extract and record these trips

3.1.3 *Results*

After preprocessing, and processing the data for errors, histograms of the extracted intercity trips are shown in Figure 3. The bars (left vertical axis) represent the percentage of trips in each travel time bin (percentages showed for privacy reasons), and the lines (right vertical axis) represent the cumulative percentage of trips in each bin.

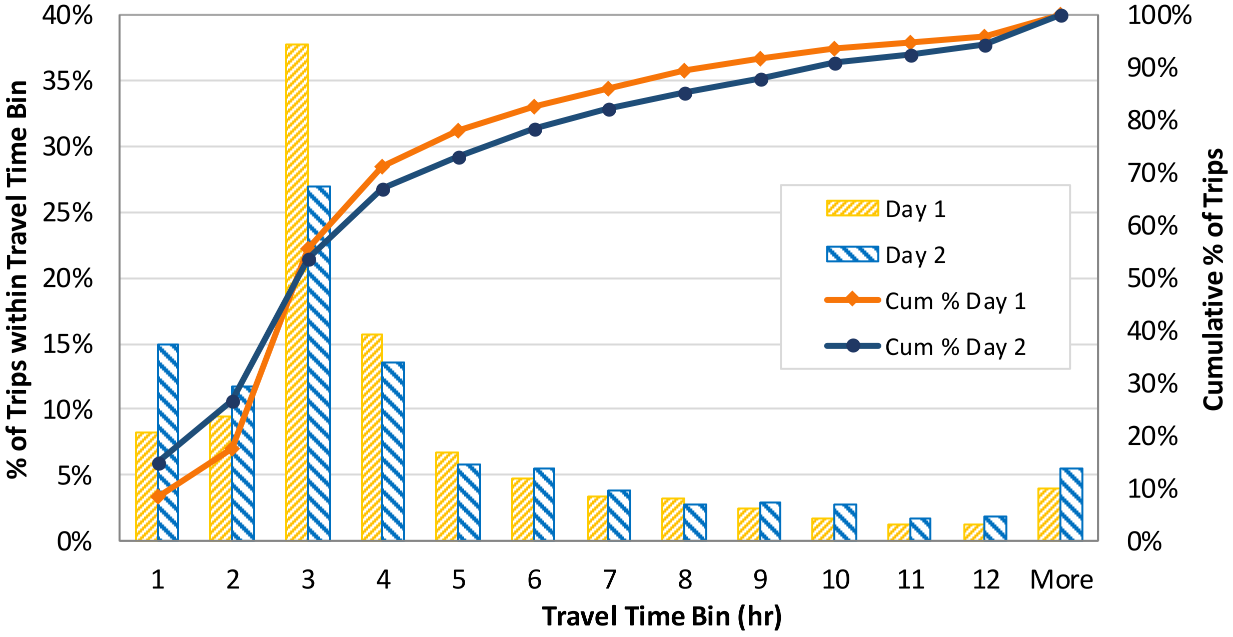


Figure 3 Histogram and cumulative percentage of intercity trips on both days.

It can be observed that the largest proportions of trips, on both days 1 and 2, are in the three-hour range, followed closely by trips in the one-, two-, and four-hour ranges. This supports the fact that most driving trips between Edmonton and Calgary take approximately three hours. As discussed previously, trips can also take longer as users may choose to make a stop and travelling by bus takes a minimum of four hours. The trips in the one-hour range are potentially air trips. The cumulative percentages and frequencies of the histogram for both days are quite similar; with Day 2 having a higher proportion of trips in the one-hour range. This may be due to a preference for flying over driving in the winter (when driving conditions can be heavily degraded and dangerous after precipitation). Descriptive statistics are shown in Table 1.

There is a fair share of trips in the sample that are longer than the four-hour range. It is reasonable to surmise that they are most likely due to stops made in between Edmonton and Calgary, such as for gasoline, meals, shopping, or other rest activities. Trips with stops made along the way are known as trip chains, however there is no formal agreement to what is considered a trip chain (ex. stop for a meal or dropping in on friends along the way) or a new trip entirely (McGuckin & Nakamoto, 2004). In this section, we will simply consider stops to belong to a trip; in the mode split analysis (Section 4.1.2) we will observe clusters of these "long" trips, or trips with stops made.

Table 1 Descriptive Statistics from Each Data Sample Day

Statistic	**Day 1**	**Day 2**
Number of Trips	1247	1297
Median Travel Time (hr)	2.8	2.8
Mean Travel Time (hr)	4.0	4.1
Standard Deviation (hr)	3.3	3.7
Coefficient of Variation	0.84	0.89
Skew	2.27	1.57

The median observed travel time for both days is about 2.8 hours, whereas the means are closer to four hours, as the histograms' long right tails are suggestive of intercity trips with stays made mid-trip. The skew for both days are positive, indicating a skewed distribution to the left as can be seen in the histogram. The skew supports the fact that intercity trips should take either less than one hour by air or between three to four hours by ground. The standard deviation is very high, this is most likely due to the nature that trips by air take much shorter than by ground, and intercity trips consisting of multiple trips and stays (thus leading to a long observed travel time) are included.

Statistical analyses are conducted to determine whether the travel times observed for each day are statistically similar. This is done to gain a better picture of the two days of data, and whether there is any consistency to the sampling. Other differences in the data could be due to seasonal variations and not from data sampling; as Day 1 occurs in the summer and Day 2 occurs in the winter. The statistical analysis seeks to see if the two data sets are distinctly similar or different.

First an f-test is conducted to compare the variances of the two data sets. Parametric tests such as the commonly used Student's t-test or f-test work on the assumption that the data is normally distributed. A test for normality was not conducted in this case, because when there is a large sample size (ie. larger than 40 samples, which is true in both data sets), the central limit theorem can be invoked, which states the sample means will be close enough to normally distributed even if the population is not (Elliot & Woodward, 2007). The null hypothesis for the f-test is that the variances of the two data sets are significantly similar at the 95% confidence interval. The results of the f-test show that the null hypothesis is rejected, further suggesting that the variances are not similar. In this case, it is not possible to use the Student's t-test to test for similarity in the two sample means, as the Student's t-test works on the assumption of equal population variances. A lesser known, but equally effective test is the Welch's t-test (Ruxton, 2006). The Welch's t-test enables the testing of two sample means with unequal variances by lowering the degrees of freedom. The null hypothesis is that the two samples means are significantly similar. The results of the Welch's t-test fail to reject the null hypothesis. A summary of the tests is provided below in Table 2.

Table 2 Parametric Analysis between Two Sample Days of Trip Travel Times

Test	**f-test**	**Welch's t-test**
Null hypothesis	The variances of the two data sets are significantly similar at the 95% confidence interval	The means of the two data sets are significantly similar at the 95% confidence interval
p-value	0.003068 < critical p-value (0.05)	0.2187 > critical p-value (0.05)
Results	Reject null hypothesis	Fail to reject null hypothesis

The f-test and the Welch's t-test indicate that although the sample means are significantly similar, the variances are not. Because the two means are statistically similar, it is possible to pool the two data sets together. Pooling the two days of data could increase the sample size, however this was not done for a few reasons. With only two days of data, it is unknown whether differences in the data (such as the significantly different variances) are from seasonal variability, data sampling variances, or differences in data errors or noise.

In conclusion, I extracted intercity trips from the anonymous cellular data sample between Edmonton and Calgary with favourable results, with a statistically similar mean travel time on both days and in the time range we would expect. The distribution of trip travel times in the 1-hr, and 3-hr range suggests that the anonymous cellular data is in fact, capable of extracting trips with reasonable time estimates. There seem to be a fair percent of trips longer than the 3 to 4-hr mark, which suggests that there may be people who are making these intercity trips with long stops in between. This will be further explored in the mode split analysis in Section 4.1.

3.2 Extracting Intercity Trips within Urban Areas of Alberta

Thus far the analysis with anonymous cellular data involves looking at intercity trips between the cities of Edmonton and Calgary. In this part of the analysis, the use of the data is expanded to all cities (the highest urban municipal status in Alberta) in the province, to extract an OD matrix of intercity trips made. Since another application of this data is identifying the mode split between air and ground, it is of value to include other urban areas in Alberta that have air services connecting with other urban areas or cities.

3.2.1 *Urban Zones Description*

According to the Municipal Government Act in the Province of Alberta, a city is defined as an area that has a population of 10,000 people or more, and there are currently eighteen municipalities in the Province of Alberta (Government of Alberta, 2016). Statistics Canada classifies an area with neighbouring municipalities around a core a census metropolitan area (CMA), with a total population of at least 100,000 (Statistics Canada, 2015). As mentioned in Section 3.1.1, Edmonton and Calgary are the only two CMAs in Alberta. In addition, the three hamlets of Fort McMurray, Fort MacKay, and Sherwood Park are included in this analysis. In Alberta, hamlets are unincorporated communities of five dwellings or more within a municipal (Government of Alberta, 2016). Sherwood Park is included due to its large population, and is part of the Edmonton CMA. Fort McMurray and Fort MacKay are areas of interest due to the high population of oil sands workers that live in Fort McMurray or commute back and forth to that area. Thus, cities, CMAs, and hamlets are included as urban zones studied in this section. Suburban cities (and hamlet Sherwood Park) within the Edmonton or Calgary CMA will not be analyzed as their own zone, as many trips between these suburban cities within the CMA are less than 40 km and not considered long-distance trips by the Travel Survey of Residents of Canada. The urban zones are shown in Table 3, with the suburban cities within each CMA specified. Their geographical location in the province of Alberta is shown in Figure 4.

Table 3 List of Urban Zones and Populations in Alberta

OD Matrix ID	Urban Zone	Classification	Population
1	Brooks	City	14,200
2	**Calgary**	**CMA**	1,469,300[1]
	Airdrie	City	14,200
	Calgary	City	1,235,200
	Chestermere	City	19,700
3	Camrose	City	18,000
4	Cold Lake	City	15,700
5	**Edmonton**	**CMA**	1,392,600[1]
	Edmonton	City	899,400
	Fort Saskatchewan	City	24,600
	Leduc	City	30,500
	Sherwood Park	Hamlet	70,618[1]
	Spruce Grove	City	33,600
	St. Albert	City	64,600
6	Fort MacKay/Oil sands area	Hamlet	N/A
7	Fort McMurray	Hamlet, Urban Service Area	66,600[2]
8	Grande Prairie	City	68,600
9	Lacombe	City	12,700
10	Lethbridge	City	96,900
11	Lloydminster	City	19,700
12	Medicine Hat	City	63,000
13	Red Deer	City	99,800
14	Wetaskiwin	City	12,600

[1](Statistics Canada, 2017b) [2](Statistics Canada, 2017a), all others: (Government of Alberta, 2017)

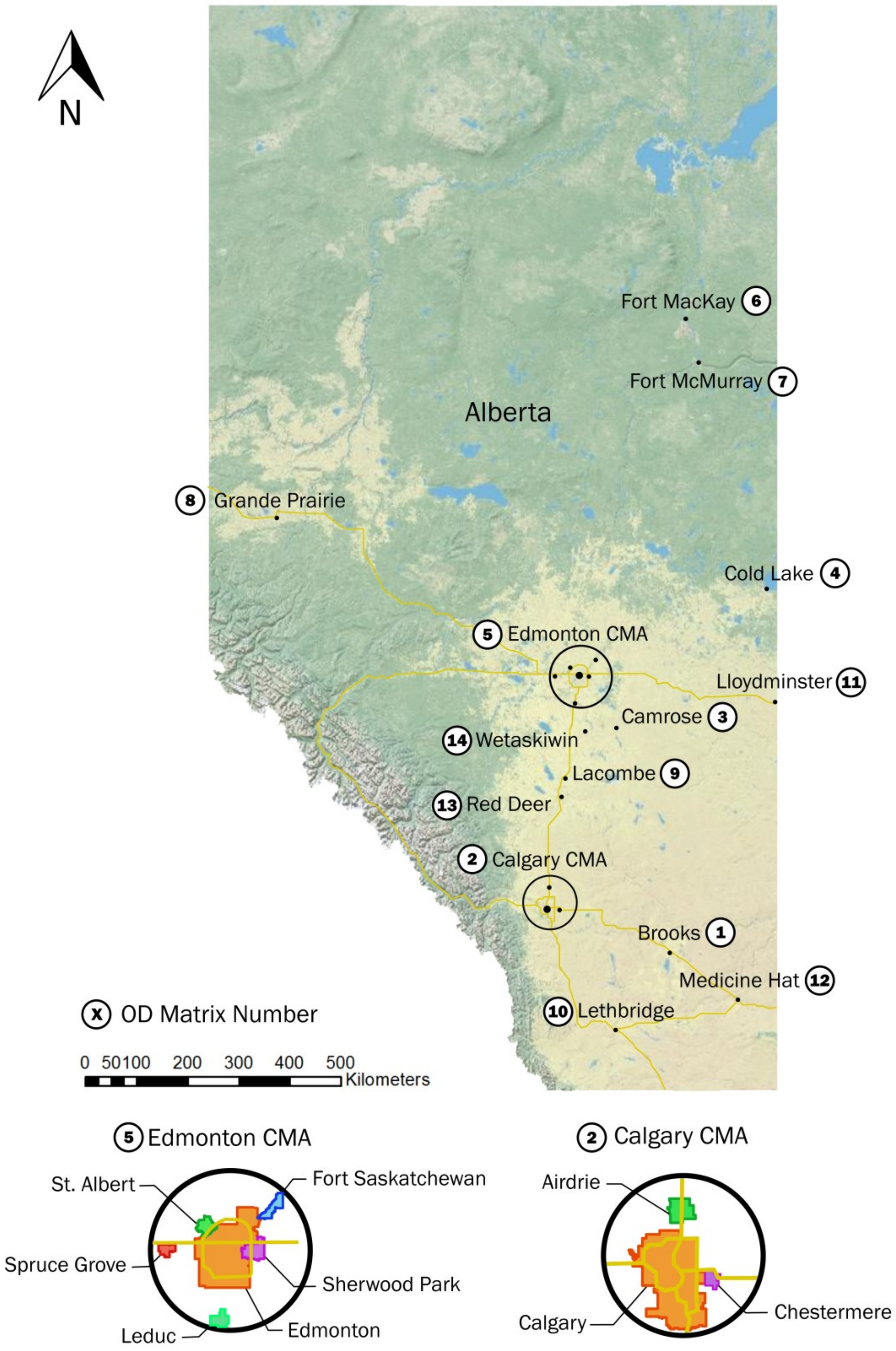


Figure 4 Map of municipalities (and other urban zones studied) in Alberta. (Base map source: (U.S. National Park Service, 2017).

3.2.2 *Methodology*

The data preprocessing involved extracting all users with records that contain at least one record in any two of the fourteen urban zones listed in Table 3. Similar to the preprocessing in Section 3.1.2, errors with the "flip-flop" effect are filtered out. The spatial resolution of each Cell and LA ID were aggregated to one of the fourteen urban zones or out of bounds. From the database, the processed and cleaned data about consisted of: Anonymous User ID, Location Zone, First Timestamp, and Last Timestamp. After preprocessing, there were 153,344 records and 47,083 unique user ID's on Day 1 and 148,096 records and 42,833 unique ID's on Day 2.

For each user, any trips they have made are extracted using the following logic:

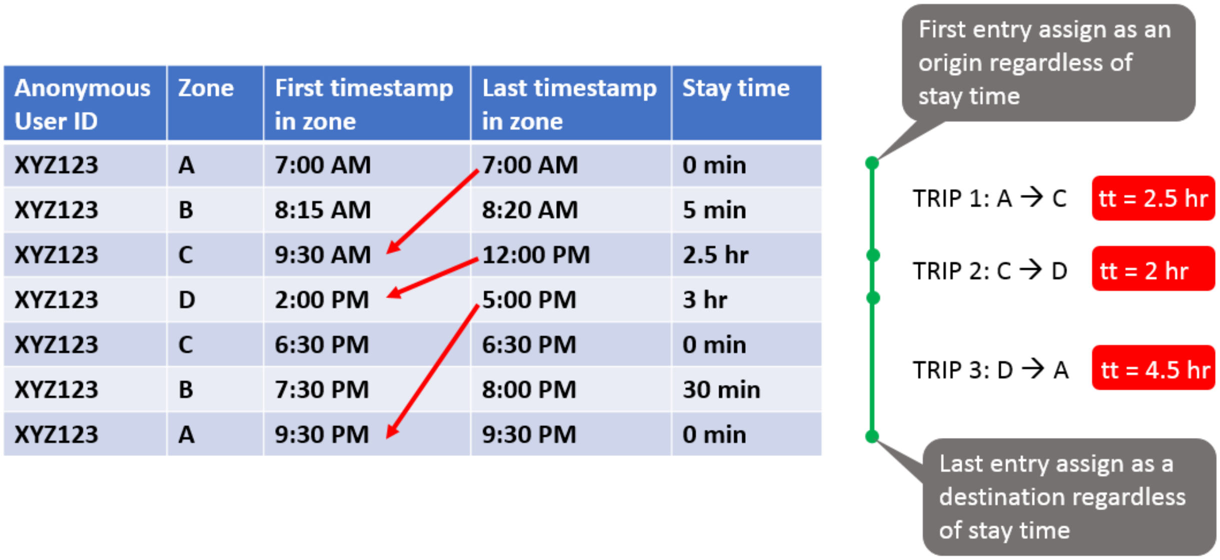
- The Stay Time for each urban zone is approximated by the time difference between the First Timestamp and Last Timestamp that occurred in that zone.
- The first entry for the user is assumed to be a trip Origin.
- The first entry following the Origin entry that has a Stay Time of one hour or greater is considered a stay. The zone of this entry is assigned as the Destination, and this trip is added to the OD matrix. The Travel Time for an intercity trip is estimated as the time difference between the First Timestamp in a Destination and the Last Timestamp in an Origin.
- This entry is then reset as an Origin, and the following entries are once again analyzed for another stay. The last entry of each user is also set as the end of a trip regardless of Stay Time.

Additionally, further data exploration led to developing additional conditions to remove invalid trips:

- If a user is out of bounds (in a zone that is not one of the fourteen zones) for over six hrs, reset Origin to next valid zone. The user may have made a trip from an urban zone to a rural area which is out of the scope.
- If a user has no records for over 12 hrs, reset Origin to next valid zone. A small percentage of users were found to have sporadic records, for example, one early morning and one late

at night. These are most likely data errors where no records were made periodically throughout the day.

Any entries (excluding the first and last) that do not have a Stay Time of at least one hour are assumed to be pass-bys. Again, there is no clear definition for what should be considered a pass-by that is part of a trip chain or a destination. A user may have just one trip (first entry as Origin, last entry as Destination), or multiple trips if there are entries in between the first and last with a Stay Time of longer than one hour. An example of this OD assignment per user is demonstrated in Figure 5, where a mock example of a user's records are shown, with three intercity trips extracted.



Anonymous User ID	Zone	First timestamp in zone	Last timestamp in zone	Stay time
XYZ123	A	7:00 AM	7:00 AM	0 min
XYZ123	B	8:15 AM	8:20 AM	5 min
XYZ123	C	9:30 AM	12:00 PM	2.5 hr
XYZ123	D	2:00 PM	5:00 PM	3 hr
XYZ123	C	6:30 PM	6:30 PM	0 min
XYZ123	B	7:30 PM	8:00 PM	30 min
XYZ123	A	9:30 PM	9:30 PM	0 min

First entry assign as an origin regardless of stay time

Last entry assign as a destination regardless of stay time

Figure 5 Illustration of anonymous cellular data records for one user and sample extracted trips.

Matlab code is then implemented on the anonymous user records that have been identified to have at least one record in at least two of the urban zones, to extract all intercity trips. Each extracted trip has an Origin, Destination, observed Travel Time, and trip Start Time. The trips are added to an OD matrix per day.

3.2.3 *Results*

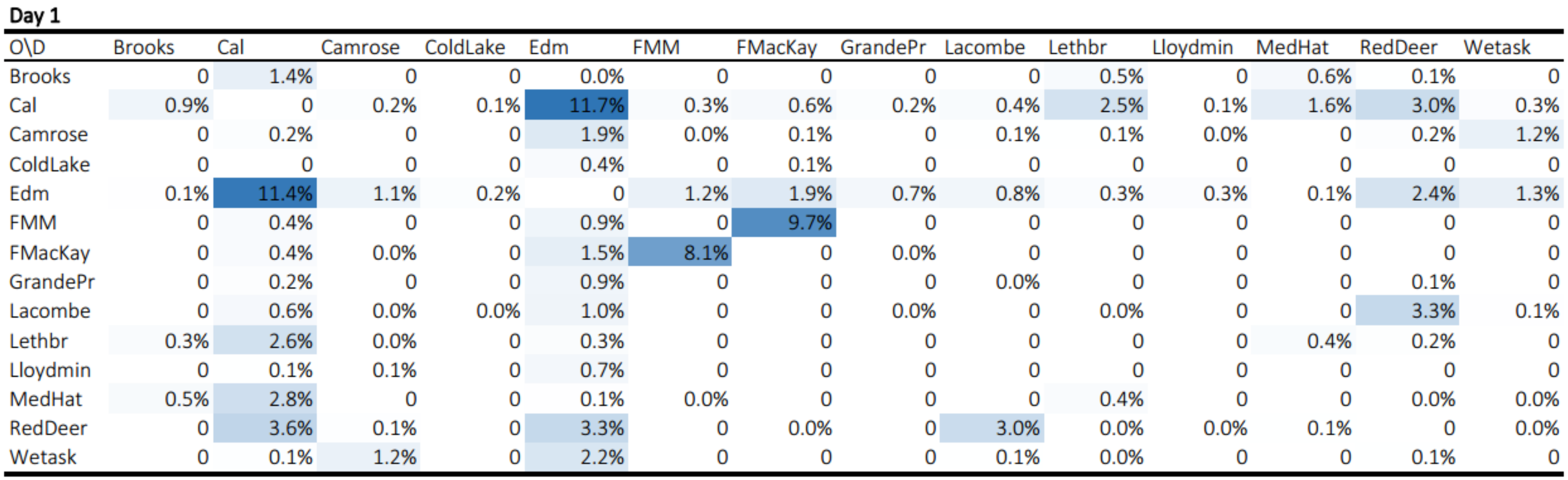
In total, 3777 trips were extracted on Day 1 and 3463 trips on Day 2. For reasons of confidentiality with small sample sizes, the number of trips between each OD pair are not shown, but a percentage derived from the total number of extracted intercity trips each day are shown in Table 4. The OD matrices for Day 1 and Day 2 are colour-coded in increasing tones categorized by each trip pair's percentile distribution. We can observe that the city pairs of Edmonton – Calgary and Fort McMurray – Fort MacKay have the highest trip volumes. As Edmonton and Calgary are the two largest cities in the province, they are likely to have the greatest travel activity between them. Fort McMurray and Fort MacKay are likely to have a relatively high volume of trips from workers commuting to and from their home in Fort McMurray and the oil sands camp sites, given the size of these employers.

There is also a fair number of trips recorded between Calgary – Red Deer, Red Deer – Lacombe, and Calgary – Lethbridge. A potential reason for more trips between Calgary and Red Deer is that Calgary is a larger city with a larger population, which naturally will generate more travel compared to another city with a smaller population (i.e. the gravity model, with populations in place of mass). As well, residents of Red Deer are more likely to use YYC than YEG due to YYC having superior air service (more direct destinations, greater frequencies, more airlines, lower airfares).

There are less trips extracted that originate or end at smaller cities, but it can be observed that the highest portion of these trips will start or end at the cities closest to them (in the gravity model, more attraction between masses at smaller distances). Wetaskiwin's largest trip pairs are with Camrose and Edmonton; Lacombe's largest trip pair is with Red Deer. Cities that are more geographically secluded from other cities, do not appear to have many trips to and from the other cities (ex. Lloydminster, Grande Prairie).

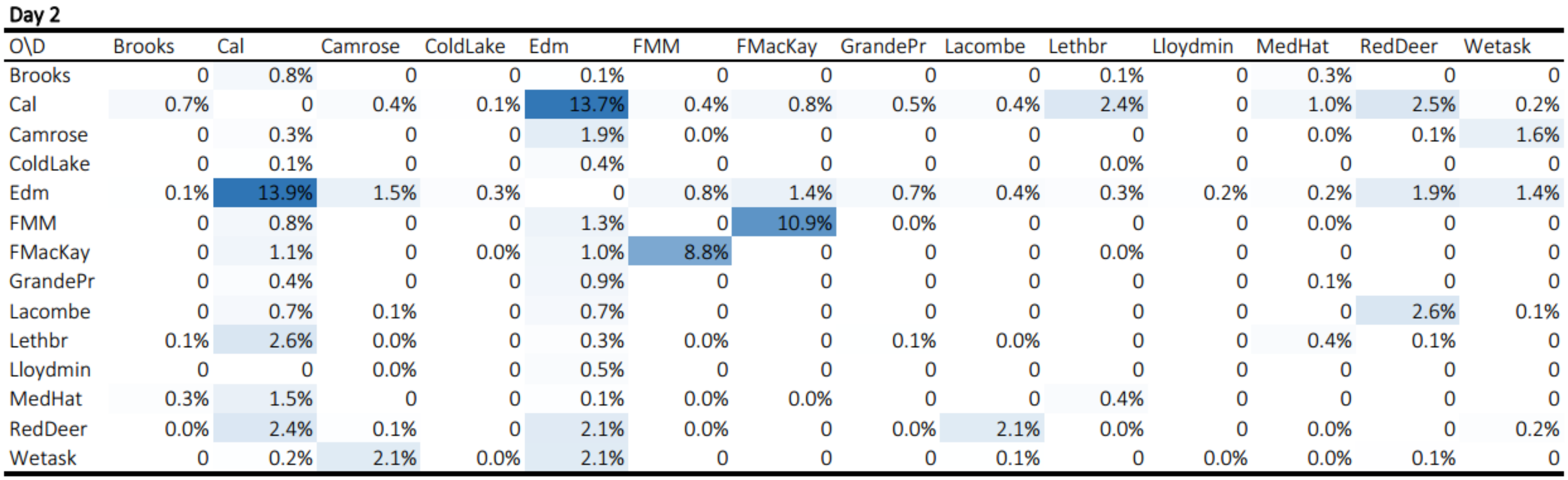
The extracted trips OD matrix confirms what one would expect: more trips will occur between larger cities (in this case Edmonton, Calgary), or areas that have high trip attraction due to employment (Fort McMurray, Fort MacKay) and trip attraction between city pairs are dependent on the distance between them (ex. Grande Prairie, Lloydminster are small cities far away with less trips; Wetaskiwin, Lacombe, has relatively more trips to cities close to them).

Table 4 Intercity Trips (%) for Urban Areas within the Province of Alberta



Day 1

O\D	Brooks	Cal	Camrose	ColdLake	Edm	FMM	FMacKay	GrandePr	Lacombe	Lethbr	Lloydmin	MedHat	RedDeer	Wetask
Brooks	0	1.4%	0	0	0.0%	0	0	0	0	0.5%	0	0.6%	0.1%	0
Cal	0.9%	0	0.2%	0.1%	11.7%	0.3%	0.6%	0.2%	0.4%	2.5%	0.1%	1.6%	3.0%	0.3%
Camrose	0	0.2%	0	0	1.9%	0.0%	0.1%	0	0.1%	0.1%	0.0%	0	0.2%	1.2%
ColdLake	0	0	0	0	0.4%	0	0.1%	0	0	0	0	0	0	0
Edm	0.1%	11.4%	1.1%	0.2%	0	1.2%	1.9%	0.7%	0.8%	0.3%	0.3%	0.1%	2.4%	1.3%
FMM	0	0.4%	0	0	0.9%	0	9.7%	0	0	0	0	0	0	0
FMacKay	0	0.4%	0.0%	0	1.5%	8.1%	0	0.0%	0	0	0	0	0	0
GrandePr	0	0.2%	0	0	0.9%	0	0	0	0.0%	0	0	0	0.1%	0
Lacombe	0	0.6%	0.0%	0.0%	1.0%	0	0	0.0%	0	0.0%	0	0	3.3%	0.1%
Lethbr	0.3%	2.6%	0.0%	0	0.3%	0	0	0	0	0	0	0.4%	0.2%	0
Lloydmin	0	0.1%	0.1%	0	0.7%	0	0	0	0	0	0	0	0	0
MedHat	0.5%	2.8%	0	0	0.1%	0.0%	0	0	0	0.4%	0	0	0.0%	0.0%
RedDeer	0	3.6%	0.1%	0	3.3%	0	0.0%	0	3.0%	0.0%	0.0%	0.1%	0	0.0%
Wetask	0	0.1%	1.2%	0	2.2%	0	0	0	0.1%	0.0%	0	0	0.1%	0



Day 2

O\D	Brooks	Cal	Camrose	ColdLake	Edm	FMM	FMacKay	GrandePr	Lacombe	Lethbr	Lloydmin	MedHat	RedDeer	Wetask
Brooks	0	0.8%	0	0	0.1%	0	0	0	0	0.1%	0	0.3%	0	0
Cal	0.7%	0	0.4%	0.1%	13.7%	0.4%	0.8%	0.5%	0.4%	2.4%	0	1.0%	2.5%	0.2%
Camrose	0	0.3%	0	0	1.9%	0.0%	0	0	0	0	0	0.0%	0.1%	1.6%
ColdLake	0	0.1%	0	0	0.4%	0	0	0	0	0.0%	0	0	0	0
Edm	0.1%	13.9%	1.5%	0.3%	0	0.8%	1.4%	0.7%	0.4%	0.3%	0.2%	0.2%	1.9%	1.4%
FMM	0	0.8%	0	0	1.3%	0	10.9%	0.0%	0	0	0	0.0%	0	0
FMacKay	0	1.1%	0	0.0%	1.0%	8.8%	0	0	0	0.0%	0	0	0	0
GrandePr	0	0.4%	0	0	0.9%	0	0	0	0	0	0	0.1%	0	0
Lacombe	0	0.7%	0.1%	0	0.7%	0	0	0	0	0	0	0	2.6%	0.1%
Lethbr	0.1%	2.6%	0.0%	0	0.3%	0.0%	0	0.1%	0.0%	0	0	0.4%	0.1%	0
Lloydmin	0	0	0.0%	0	0.5%	0	0	0	0	0	0	0	0	0
MedHat	0.3%	1.5%	0	0	0.1%	0.0%	0.0%	0	0	0.4%	0	0	0	0
RedDeer	0.0%	2.4%	0.1%	0	2.1%	0.0%	0	0.0%	2.1%	0.0%	0	0.0%	0	0.2%
Wetask	0	0.2%	2.1%	0.0%	2.1%	0	0	0	0.1%	0	0.0%	0.0%	0.1%	0

4. INFERRING TRIP MODE

One potentially important use of anonymous cellular data is to infer the mode split of OD flows. This could be particularly useful in cases where corridor service improvements or new modes are being considered for major public (or in some cases, private) investment. In this chapter the trip mode for extracted trips between Edmonton and Calgary (from Section 3.1) are found using two methods: classification and hierarchical clustering. Historical flight schedules between Edmonton and Calgary are provided from Edmonton Airports (a not-for-profit corporation that manages YEG) and used to compare with the inferred air trips on the two days. Then, the trip mode for all trip pairs in Alberta (between pairs that also have air service) is inferred using the travel time by hierarchical clustering.

4.1 Inferring Trip Mode Between Edmonton and Calgary

Two different methods were used to infer the trip mode by air or by ground. The first method involves simple categorization by travel time which were determined by logic and judgement, and was described in Hui, Wang, Kim, & Qiu (2017). Strict lower and upper travel time bounds are set for ground and air trips, and all trips with travel times falling between each of those two bounds are assigned that mode. This classification method is simple and easy to use; however, the upper and lower bounds are set arbitrarily on what we think is reasonable. The second method uses a hierarchical clustering scheme which identifies estimated travel time clusters. These clusters may then represent air trips, ground trips, or longer ground trips with stops. The advantage to clustering is that each group or cluster is determined mathematically based on the data alone, and not on an a-priori assumption or decision; points that are closest together are grouped together. Clustering requires the analyst to interpret if and what each cluster represents, and to select the appropriate number of clusters (for example the analyst may already have an idea of how many clusters there should be).

4.1.1 *Method 1: Categorize travel times*

The average travel times between Edmonton and Calgary was used to determine a reasonable upper and lower bound to categorize whether an intercity trip was made by air or ground. For air trips, people are most likely to turn their phone off before takeoff, for example before boarding the aircraft or when they are notified to do so while taxiing to the runway. If a user did not turn off their phone during the flight, there could still be a signal picked up during the ascent and descent phases when the aircraft is at lower altitudes. To account for these various behaviors and circumstances, I reasoned that 30-90 minutes is an acceptable time range for a trip identified to be made by air in the cell phone data. These upper and lower bounds encompass the average flight time of 43.1 minutes (calculated from historical flight data provided by Edmonton Airports).

For ground trips, the approximate driving time between Edmonton and Calgary is about three hours; it is of course highly dependent on where a user's trip is starting from and ending at (city centre, north end, south end, etc.) within each city. In addition, a trip made by vehicle can be much longer than three hours when stops are made in between, or there is heavy traffic congestion (likely through the cities – Edmonton, Calgary, Red Deer). As mentioned, one of the key limitations of using passive data such as anonymous cellular data to identify trips is that there is no knowledge about a trip maker's purpose and other characteristics. At first, the binary Between variable was used to identify air trips; if there were no records in between city 1 and city 2, then a trip was assigned as air, otherwise it was assigned as ground. However, this was found to capture too few trips, leaving a large number of short trips unclassified. The method of using this Between variable was determined to be infeasible in this study given the limited data.

Instead, the travel time was used to categorize the mode of a trip with the following conditions:

- If the trip travel time is between 0.5 – 1.5 hr → Assign as an air trip
- If the trip travel time is between 2 – 6 hr → Assign as a ground trip
- Otherwise, trip is unclassified

There were quite a few trips longer than 8 hours; one possible reason for this is that the user made multiple trips with stays in between. These trips are considered to be out of our scope. Figure 6 represents each intercity trip by a dot, plotted along the x-axis by the trip start time with the trip length shown on the y-axis. The plot on the left is for Day 1, and Day 2 on the right.

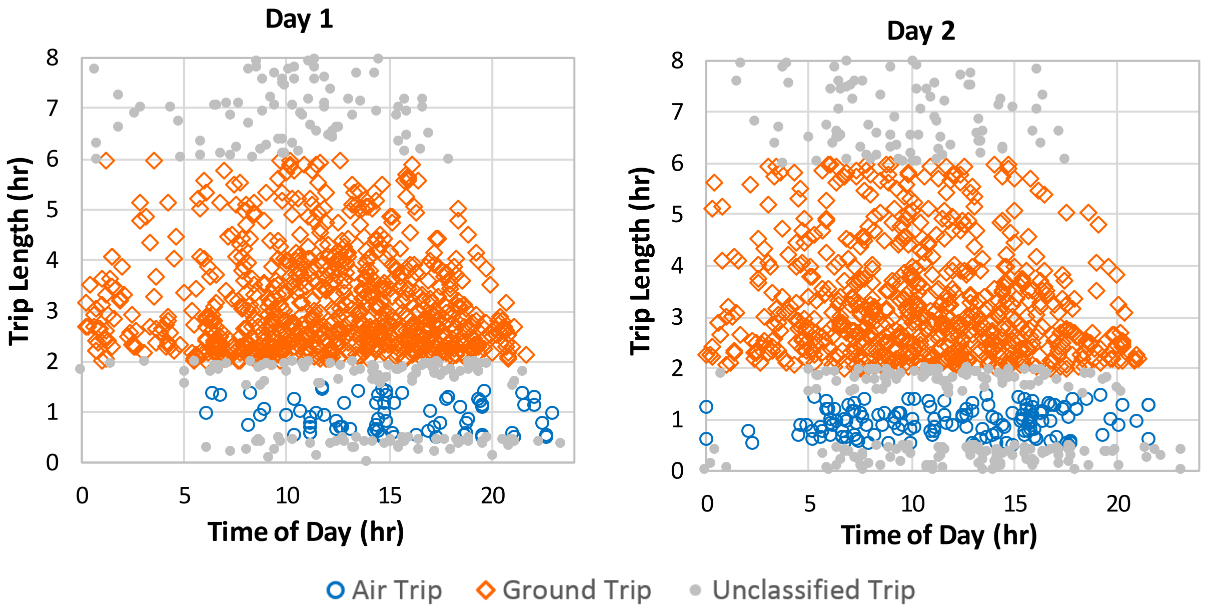


Figure 6 Intercity trip plots with trip mode classified. Taken from "Investigating the use of anonymous cellular phone data to determine intercity travel volumes and modes" by Hui et al, 2017.

From Figure 6, it appears that air trips do not begin until around 6 am while as ground trips occur throughout the day. This is because air trips are dependent on flight schedules whereas drivers can choose to conduct their trip anytime. The limitations of this mode classification can also be observed in Figure 6 by the unclassified trips just underneath the 2-hr mark or just under the half hour mark. These unclassified trips are very close to their respective group of ground and air trips, but was not classified due to what the limiting bounds are (which are determined arbitrarily). The mode split as determined from this classification are shown in Table 5 for each day, along with descriptive statistics.

Table 5 Intercity Trip Mode Split and Travel Time Statistics

	Day 1		**Day 2**	
Direction	**Air**	**Ground**	**Air**	**Ground**
Southbound	11%	89%	18%	82%
Northbound	7%	93%	19%	81%
Total	9%	91%	18%	82%
Travel Time (hr)				
Mean	0.94	3.12	0.96	3.27
Standard Deviation	0.31	0.97	0.27	1.09

The mode splits found are reasonably close to estimates from an Edmonton-Calgary corridor HSR study, which found 10% by air and 90% by ground (which is further broken down to 86% by auto, 4% by freight trucks) (TEMS, Inc. and Oliver Wyman, 2008). The higher ratio of air trips picked up on Day 2 than Day 1 may be due a seasonal preference for flying over driving in the winter, or sampling variances. For example, it appears there is more noise in the Day 2 data compared to Day 1, which can be seen by the categorized flight trips to have occurred at the beginning of the day when there were no flights scheduled. Because of the limited days of data, it is hard to determine to what extent the sampling varies, and as the f-test in Section 3.1.3 has shown, the variances for the two days (for all trips, not just classified as air or ground) are significantly different.

Overall, this method seems to provide a reasonable estimate of the mode splits between air and ground. However, the limits set a-priori leave a large number of trips unclassified - for example, the trips between classified air and ground trips. As well, the large number of trips that are longer than the expected driving time (~3 hrs) suggest that there may be another group that represents people making trips and taking stops along the way or trips by bus. In the next section, a hierarchical clustering application is introduced, to identify how many groups naturally form in the data.

4.1.2 *Method 2: Clustering of Travel Times*

The primary purpose of data clustering is to group a given dataset into groups where the data points within each group are as similar as possible (Aggarwal & Reddy, 2014). A common method of clustering is to use distance-based algorithms. The two primary types of clustering are hierarchical clustering and partitional clustering (Kaski, 1997). K-means clustering (the most common form of partitional clustering) and hierarchical clustering were applied on the data. The k-means clustering result appeared inferior to the results of hierarchical clustering, as the trips in each cluster did not associate as well with each other upon different iterations (for example trips that looked part of the driving trips cluster would be assigned as part of the flight trip cluster).

The algorithm for hierarchical clustering schemes (HCS) was developed in 1967, and it is based on the distances between a data point and all other data points (Johnson, 1967). In HCS, clusters are either merged from the original data set until there is one cluster (bottoms up/agglomerative) or split from one large cluster into smaller ones (top down/divisive approach). An agglomerative HCS algorithm was used with the following steps:

1. Each travel time observation (i.e. data point) begins as its own cluster. With X data points, there are X clusters initially.
2. Distances between each cluster are calculated, which are known as linkages. There are different ways to calculate the distances between clusters when there is more than one data point within each cluster. Common methods include single-linkage (the shortest distance), average-linkage (average distance between links), complete-linkage (the longest distance). In this case, average-linkage was used.
3. The two closest clusters based on the linkages are merged into one cluster. Now there is one less cluster.
4. Repeat steps 2 and 3 until there is only one cluster left.

As seen in the algorithm, every data point starts as its own cluster and is slowly grouped together until the whole dataset is in one large cluster. An important feature of hierarchical clustering is that the number of clusters is not decided a-priori before clustering. Although we are looking for at least two clusters (one representing air trips, the second ground), from Section 4.1.1 it appears there is the possibility of longer drive trips (potentially with stops) which may form their own cluster as well. A dendrogram, or tree diagram presents a visualization of the hierarchical clustering (see Figure 7). Each data point begins as a single node on the x-axis, and the merging of new clusters are represented by branches that merge into new nodes as each branch extends upwards. The y-axis of the dendrogram represents the linkage distance between clusters (which in this analysis is measured in time, since each data point is a travel time). For example two branches that merge at approximately 2.1 hr circled in red in Figure 7 indicates that the average linkage distance between those two clusters were 2.1 hrs. On that iteration, those two clusters had the shortest distance and thus were merged. Figure 7 is the dendrogram of the extracted intercity trips less than 8 hours from Day 1.

The number of branches when a dendrogram is “cut” at a certain y-axis distance will be the number of clusters. For example, cutting the dendrogram at 3 hr on the y-axis will have two branches, representing two clusters. They merge at approximately 3.6 hr on the y-axis, indicating that the average travel time difference between the two clusters is 3.6 hrs. The dendrogram lets the analyst observe if and where there are large partitions in the data, which is represented by nodes that merge a lot higher from the nodes below it. There are also various criteria and methods that can be used to find the optimal number of clusters in a data set. In this analysis, we are looking for at least two clusters, to represent air and ground trips. A package in R called NbClust developed by Charrad, Chazzali, Boiteau, & Niknafs (2014) was also used to identify what 26 different clustering criteria and indices determined to be the optimal number of clusters in the data. The code and results are shown in Appendix A, but the results from all criteria/indices did not provide a large majority for a certain number of clusters to be optimal. Here I determine the number of clusters by identifying how many are needed to show a differentiation between air and ground trips, as well as by looking at natural partitions in the dendrogram.

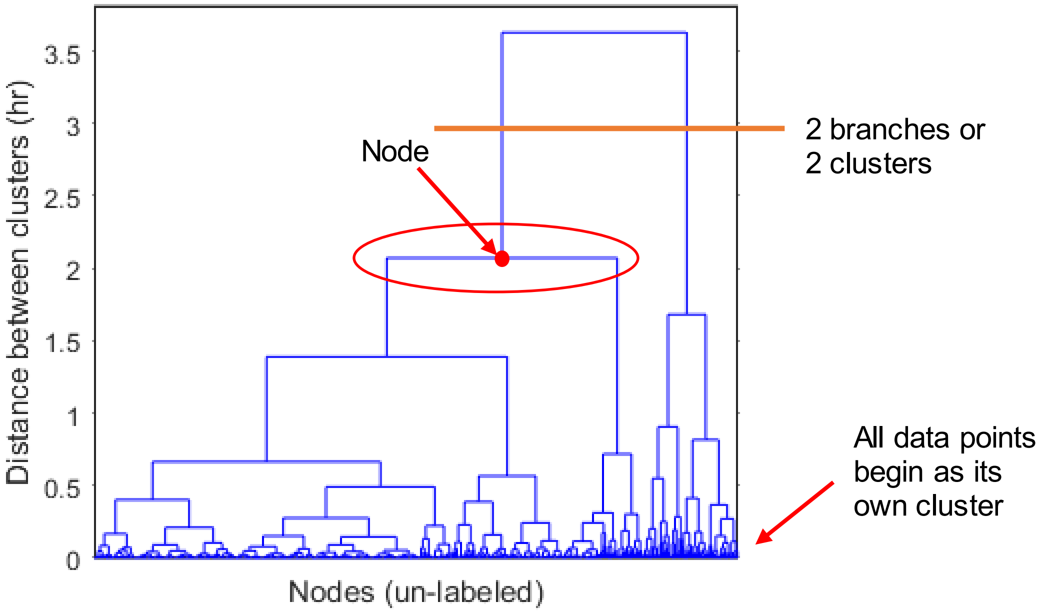


Figure 7 Dendrogram of Day 1 Intercity Trips less than 8 hrs.

From observing the dendrograms, three travel time clusters clusters were chosen for both days. In Figure 8, the dendrograms are plotted on the left and on the right is a plot of each trip as a data point where the x-axis shows the start time of the trip and y-axis the travel time. The data points are coloured based on cluster membership. Table 6 shows the mean, standard deviation, coefficient of variation, and minimum and maximum observation for each cluster.

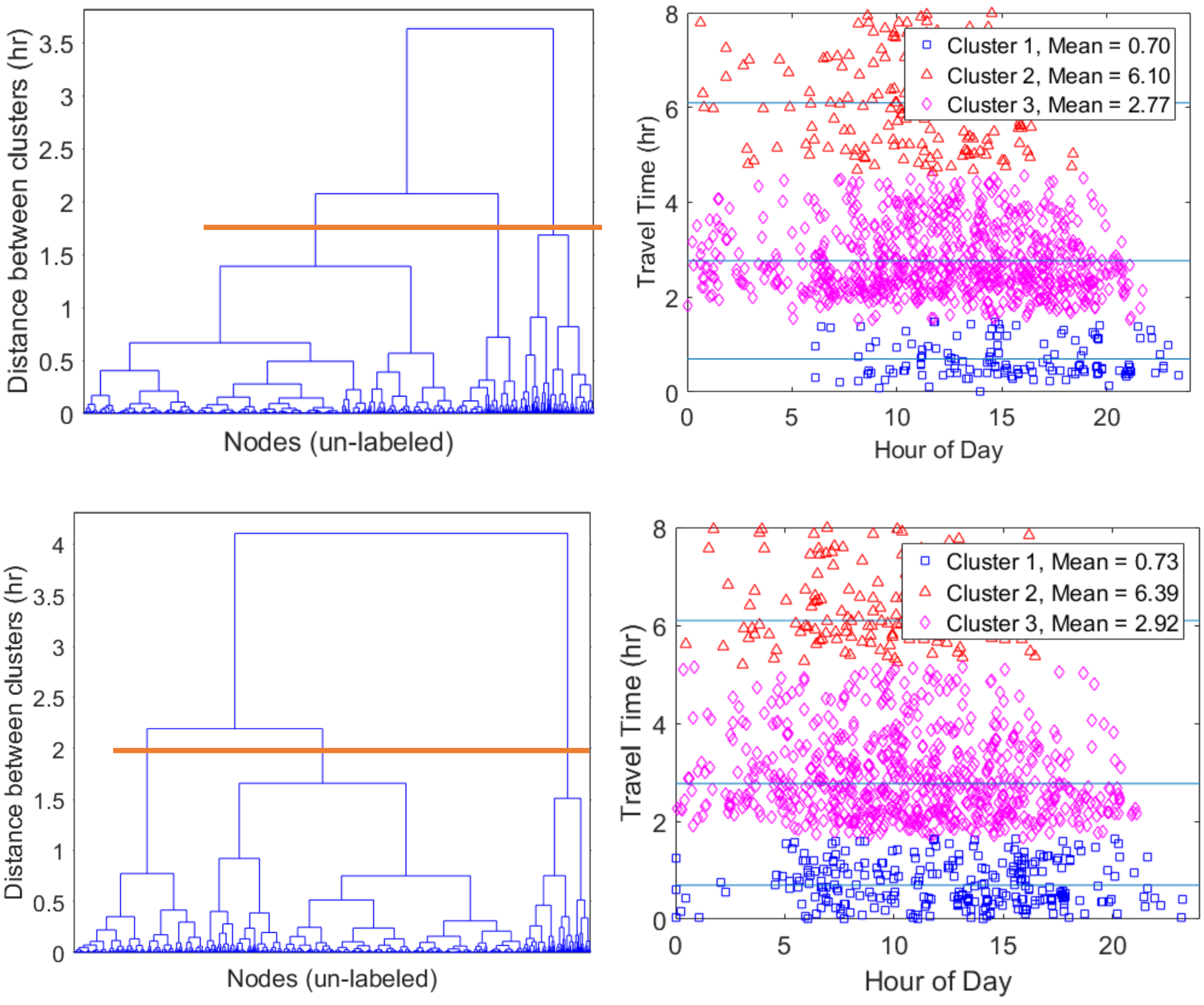


Figure 8 Dendrogram (left) and clustered trips (right) between Edmonton and Calgary (top: Day 1, bottom: Day 2)

Table 6 Descriptive Statistics of Clusters for Trips Between Edmonton and Calgary

Statistic	**Mean (hr)**	**Standard Deviation (hr)**	**Variance (hr^2)**	**Minimum (hr)**	**Maximum (hr)**
Day 1 Mode/Cluster					
Air	0.70	0.38	0.14	0.02	1.48
Ground (no stops)	2.77	0.70	0.50	1.53	4.56
Ground (with stops)	6.10	0.96	0.92	4.63	7.98
Day 2 Mode/Cluster					
Air	0.73	0.44	0.20	0.01	1.64
Ground (no stops)	2.92	0.85	0.73	1.67	5.15
Ground (with stops)	6.39	0.77	0.59	5.20	7.99

From Figure 8 there is a clear distinction of a cluster of short trips in blue, and the next cluster in pink with the trips occurring densely with a travel time between 2 – 3 hrs. The cluster means of 0.70 hr/0.73 (blue) and 2.77 hr/2.92 hr (pink) are a close match to actual average air (43.1 minutes or 0.72 hrs) and ground travel times (~ 3 hrs) respectively. As well, the earliest trip in the potential air trip cluster for Day 1 is at 6:07 am, which coincide with the scheduled flights that began at this time of the day. The cluster of the longer trips (red), with means over 6 hrs, are most likely drive trips with stops made. However the portion of trips that have errors in their travel time is unknown; for example on Day 2 there are some trips clustered as air with start times earlier than actual flights. From Table 6 we see that the clusters representing air trips on both days have the smallest variances. However, for Day 2 the ground (no stops) cluster has a higher variance than the ground (with stops) cluster. This is confirmed in Figure 8; the ground (no stops) cluster is much more spread out on Day 2 than on Day 1, which may be indicative of more data noise or errors on Day 2. The maximum observations in the ground with stops cluster show that only trips with travel time of 8 hr and less were included in this mode split analysis.

Table 7 Mode Split Summary from the Two Methods

	Method 1: Classification		**Method 2: Clustering**	
Mode/Cluster	**Day 1**	**Day 2**	**Day 1**	**Day 2**
Air	**9%**	**18%**	**12.4%**	**25.2%**
Ground (no stops)			72.8%	61.6%
Ground (with stops)			14.8%	13.2%
Ground Total	**91%**	**82%**	**87.6%**	**74.8%**

Table 7 provides the mode split summary from the two days for each method. The two mode inference methods provide somewhat similar results. Both methods found that Day 2 had approximately double the percentage of air trips than on Day 1. Throughout the analysis, there appears to be more noise on Day 2; the clusters between air and ground on Day 2 are less distinct from each other. The clustering method results in a higher share of air trips then the classification method, this may be because there was a narrow 1-hr window for trips to be classified as an air trip, and many potential air trips were unclassified. The majority of the ground trips were categorized as direct trips with no stops.

The biggest limitation with Method 1 is that there is no mathematical analysis that supports the selected bounds used to categorize each trip. The bounds used could be too limiting or too loose, for example there is not enough information about the data sampling that can explain whether an observed 29-minute trip between Edmonton Calgary is a sampling error or in fact a trip. The clustering technique is much more powerful as it can put every trip into a cluster based on the trip travel time's "distance" to other travel times. Furthermore, it can identify and distinguish how many clusters there are occurring in the data; whereas with Method 1 the analyst has no criteria or dendrogram to decide the appropriate number of groups. For these reasons, only the clustering method will be used for the trip inference analysis for all intercity trips in Alberta (Section 4.2).

4.1.3 ***Flight Times with Flight Trips Analysis***

With the historical flight data provided by Edmonton Airports, an assessment of whether the inferred flight trips could be associated with the actual flight times is performed. As described earlier, this data provided the actual takeoff and landing times of each flight between the two cities on the two sample days. From this data, there were 50 flights, 2840 filled seats on Day 1 and 58 flights, 2783 filled seats on Day 2 in both directions between YEG and YYC.

Earlier in this paper, the travel time of each trip was used to infer the trip mode through two methods, of which clustering was determined to be advantageous. To compare whether the inferred flight trips matched up with the actual flight schedules, if the midpoint of the trip fell between the takeoff and landing time of a flight, then this trip was considered a match. This criterion was chosen because it does not require conditions on the trip start and end time, simply because we do not really know when a user may turn off their phone before takeoff and turn it back on after landing. Plots showing this matching analysis is shown in Figure 9 (Day 1) and Figure 10 (Day 2) for trips in both directions. Clustered flight trips are sorted chronologically by their start times and plotted as a horizontal line with the start and end times represented by their horizontal position on the figure. Greyed out areas represent an actual flight. A trip with a black line represents a match and blue not a match; the percentage match for each travel direction and day are summarized in Table 8. Ideally, if the anonymous cellular data were to consist of travelers turning their phones on and off before and after a flight at some required time, with 100% compliance, we would see horizontal black lines stacked together with similar start and end times to their lines over the timeframe for each flight (perhaps similar to that circled in red in **Figure 9**, top right hand corner).

However this is not the case for two reasons; first, flight start and end times overlap each other during the day, making it harder to see distinct groups of users for each flight. Secondly, as the trips plotted are from the air cluster, there are errors and trips that may not actually be flight trips included.

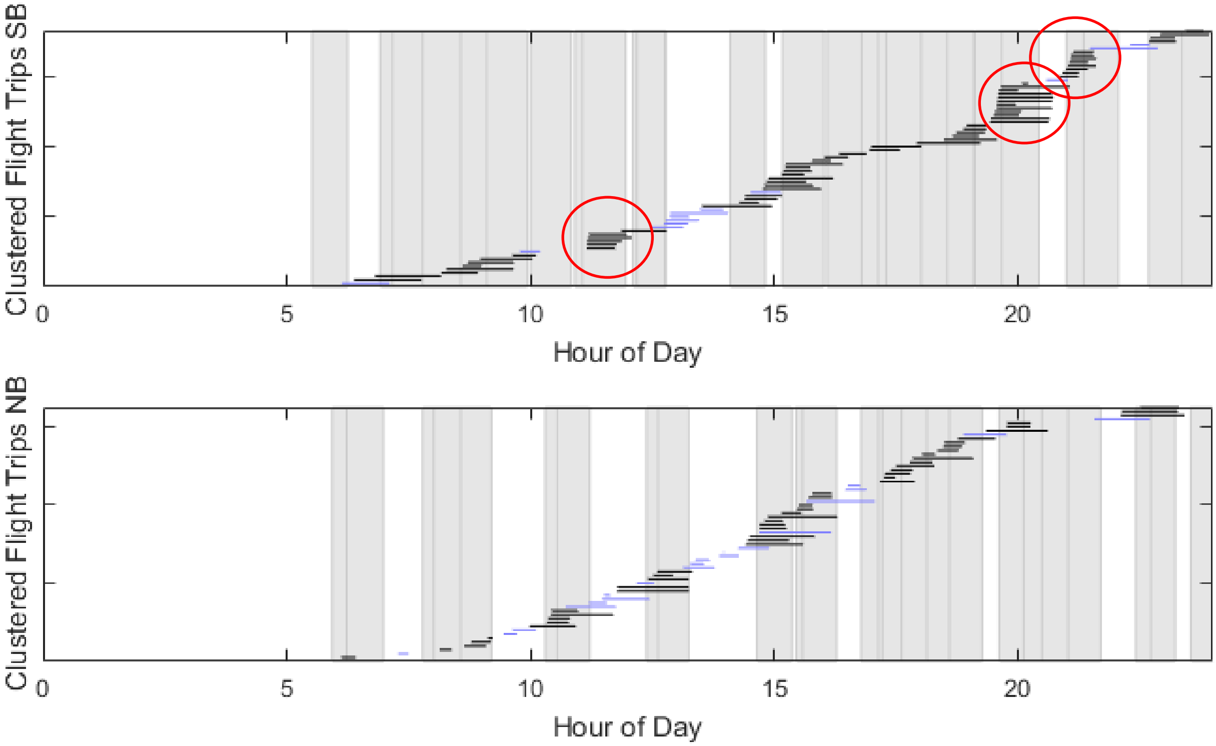


Figure 9 Day 1 clustered flight trips matched with actual flights (top: SB, bottom: NB).

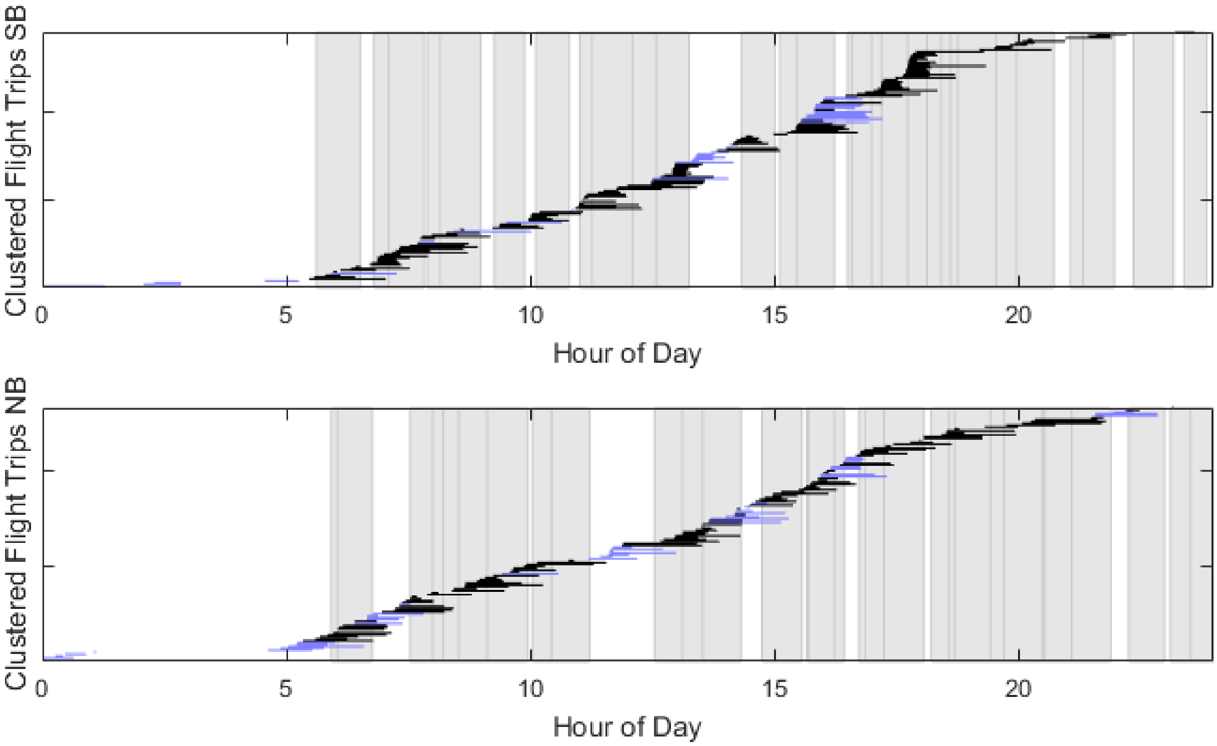


Figure 10 Day 2 clustered flight trips matched with actual flights (top: SB, bottom: NB).

Table 8 Air Trip Matches with Flight Data

Day	**% Matched Air Trips from Midpoint Criteria**
Day 1 SB (YEG-YYC)	84%
Day 1 NB (YYC-YEG)	69%
Day 2 SB	79%
Day 2 NB	69%

The results can be seen in Table 8, with successfully associated trips termed “Matched Air Trips”. Accounting for charter flights that were not tracked (only commercial flights were), the percentage of successfully associated trips are not very high and it is uncertain whether it is all due to sampling errors or the behaviour of users turning their phones on and off. However it gives confidence that with a better data sample this analysis can prove to be useful. Conducting this analysis with a full dataset will increase the penetration rate and increase the percentage of successfully associated trips. In the future, another clustering technique may be used in this analysis by clustering the inferred flight trips around the average flight times, however the dataset is too small in this case to conduct effectively.

4.2 Inferring trip mode between all intercity trip pairs

In Section 4.1, the mode split of intercity trips between Edmonton and Calgary were inferred using two methods, and I will apply the second of the two (clustering) to assess mode split for intercity trips in Alberta that have air service between them that was extracted from the anonymous cellular data set sample. Using the air service schedules of Alberta airports, ten intercity trip pairs were identified to have air service between them. Fort MacKay does not have a public airport, but there are three private airstrips owned by different oil sands companies. These include the Fort MacKay/Firebag Aerodrome (YFI), operated by Suncor, the Fort MacKay/Horizon Aerodrome (YNR), operated by Canadian Natural Resources Limited, and Fort MacKay/Albian Aerodrome (JHL), operated by Shell. The ten pairs are listed below in Table 9 with average travel times found via Google Maps for ground trips and Google Flights for air trips.

Table 9 Estimated Travel Times of Intercity Trip Pairs with Possible Air Service

Trip Pair Number	Trip Pair	IATA Airport Code	Average Travel Time (min)		Ratio of Air over Ground Travel Time
			Ground	Air	
1	Edmonton/Calgary	YEG/YYC	180	50	0.28
2	Edmonton/Grande Prairie	YEG/YGU	270	68	0.25
3	Edmonton/Fort McMurray	YEG/YMM	240	64	0.27
4	Edmonton/Fort MacKay	YEG/YFI/YNR/JHL	300	64	0.21
5	Calgary/Grande Prairie	YYC/YGU	420	90	0.21
6	Calgary/Lethbridge	YYC/YQL	120	48	0.40
7	Calgary/Fort McMurray	YYC/YMM	410	90	0.22
8	Calgary/Fort MacKay	YYC/CFG6	450	90	0.20
9	Calgary/Red Deer	YYC/YQF	80	40	0.50
10	Calgary/Lloydminster	YYC/YLL	300	70	0.23
				Average:	0.28

The air/ground travel time ratios will help infer what mode each cluster represents later in this section. However, it is important to note that there is not a standard average ratio of an average flight time to the average ground travel time range, even though the ratios in Table 9 are mostly within the 20 – 30% range. For these ten pairs, the flight ratios are mostly in similar ranges because these flights are all within Alberta and relatively short. This ratio will increase as the distance between an origin and destination decrease, though not linearly. This is because a flight has three main phases in the air: the initial climb, en route, and approach (descent) phase (Commercial Aviation Safety Team, 2013). The climb and descent phase will take 20 – 30 minutes regardless of the flight distance. This means that shorter flights will have a higher proportion of time spent on the initial climb and the approach phase. With lengthier flights, the air travel time will be significantly shorter than ground travel time. As well, average flight times are often shorter than their scheduled time, due to a practice by air carriers called "schedule padding." Schedule padding adds extra buffer time to a flight's gate-to-gate time, with the purpose of improving a carrier's reliability and on-time statistics (Skaltsas, 2011). To summarize, since all flight pairs considered in this analysis are within Alberta and relatively short, all ratios of air travel time over ground travel time are similar.

In this analysis, trips in both directions for each trip pair are included (ex. Edmonton to Calgary and Calgary to Edmonton). Once again the sampled number of trips are not provided, but the percentages are. Similar to Section 4.1.2, hierarchical clustering is applied to identify clusters of trips that may represent different modes (air, ground with no stops, ground with stops). As there are ten trip pairs each with different travel times, a unitless, rescaled travel time is introduced so that all extracted trips belonging to these ten trip pairs can be assessed as one dataset.

The travel time is rescaled by dividing the observed travel time of a trip by the average travel time specific to that trip pair, which results in a unitless travel time. However, rescaling the travel times by the average flight time did not create favourable clusters, thus the approach used was to rescale by the average ground travel time. Appendix B shows the clustering from the average flight times. Rescaling by the average ground travel time created clusters where the separation of the smallest two clusters matched the natural division seen visually (i.e. between data points around 0.2-0.3 and 1). The formula for rescaling the travel time is shown below:

$$Travel\ Time_g^i = \frac{Actual\ Travel\ Time_x^i}{Average\ Ground\ Travel\ Time_x}$$

Where

The superscript $i = 1 \dots n$ denotes the observed trip number.

The subscript g denotes the travel time rescaled by the average ground travel time.

The subscript x indicates the trip pair number that the trip observation belongs to.

$Travel\ Time_g^i$ is the unitless rescaled travel time.

Clustering using all the trip observations from the ten intercity pairs proved challenging to observe distinct clusters between air and ground trips, as there were too many very long trips that interfered. To better identify whether there was a distinction between air and ground trips, the observations were filtered to include only trips that had a $Travel\ Time_g^i$ less than three. In other words, trips that were longer than three times the average ground travel time were not included. Similar to the clustering conducted in Section 4.1.2, three clusters are chosen to represent trip modes: air, ground with no stops, and ground with stops. For the Day 2 data, from the partitioning of the data (can be seen in the dendrogram), four clusters were identified to see the air trip cluster (which is found to be optimal from criteria/indices in Appendix A). From Figure 11, we see that the cluster with the smallest $Travel\ Time_g$ has means of 0.22 and 0.33 on Day 1 and Day 2, respectively. This cluster then represents air trips, as from Table 9 the average travel time ratio of air over ground is 0.28, and this cluster is well within that range. Then the second cluster with mean of 1.06 (Day 1) and 1.10 (Day 2) represents direct drive trips, and the third cluster (Day 1 mean: 2.21, Day 2 mean: 2.12) represents drive trips with stops. Since a cutoff time of $Travel\ Time_g < 3$ was used for this analysis, the trips falling in the fourth cluster on Day 2 with mean 2.89 can be seen as outliers representing even longer trips. In the mode split table, these trips in the fourth cluster are aggregated with trips in cluster three.

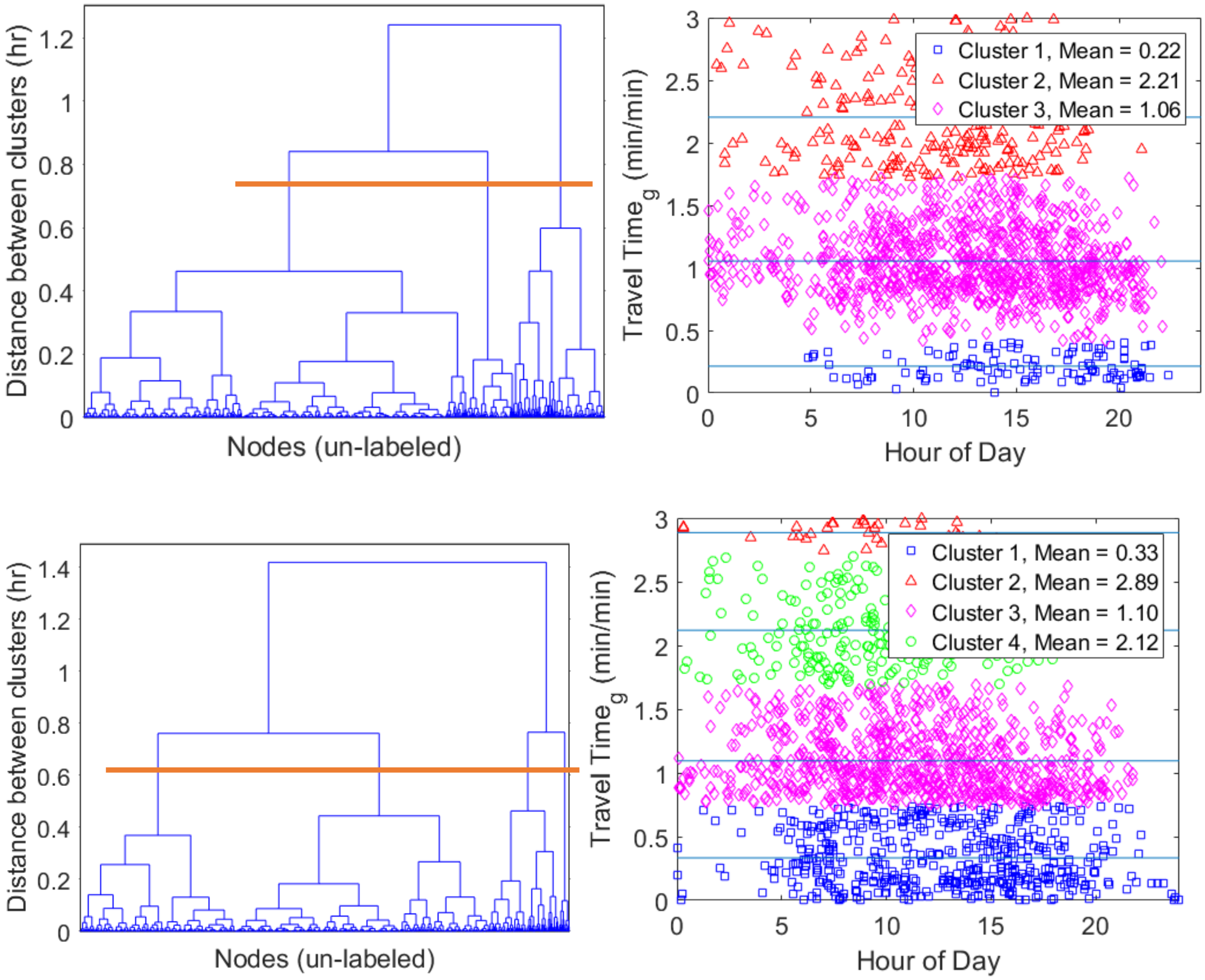


Figure 11 Dendrogram (left) and clustered trips (right) for ten intercity trips pairs rescaled by average ground travel time (top: Day 1, bottom: Day 2)

As indicated in Section 4.1.2 and 3.1.3, there may be more noise and the clusters between the air and direct ground trips less distinct on Day 2 (Figure 11). The air cluster also had a higher mean of 0.33 compared with 0.22, though both are still similar to the average ratio of 0.28. As well, based on Day 2's dendrogram, the data (all trips less than three times the average ground travel time) needed to be clustered into four groups, resulting in a fourth cluster in red. This suggests that there were a small number of trips that are almost three times the length of an average driving trip that are closely associated with each other. Descriptive statistics for each cluster are detailed in Table 10. When each clustered trip was categorized into each trip pair, the results of the clustering are shown in percentages in Table 11. The actual volume of trips sampled are very small and as a result, not shown for privacy reasons. Recall that all values are unitless.

Table 10 Descriptive Statistics for each Cluster (intercity trips across Alberta)

Statistic	Mean	Standard Deviation	Variance	Minimum	Maximum
Day 1 Mode/Cluster					
Air	0.22	0.10	0.01	0.01	0.10
Ground (no stops)	1.06	0.28	0.08	0.42	1.72
Ground (with stops)	2.21	0.35	0.12	1.73	2.99
Day 2 Mode/Cluster					
Air	0.33	0.22	0.05	0.00	0.74
Ground (no stops)	1.10	0.25	0.06	0.74	1.69
Ground (with stops)	2.23	0.37	0.14	1.70	3.00

The descriptive statistics in Table 10 confirm the results observed in Table 6 for clusters of trips between Edmonton and Calgary. In Table 10, the fourth cluster of longest trips on Day 2 is grouped together with the cluster “ground (with stops).” The variance is the highest for the ground (with stops) cluster of trips and lowest for air trips. This is because the air and ground (no stops) clusters are more distinctly clustered together while the ground (with stops) trips range from around 1.7 hr/hr to 3 hr/hr. The maximum observations in the last cluster (ground with stops) are around 3 hr/hr, as only trips with 3 hr/hr and less rescaled travel time were included in the clustering process.

Table 11 Mode Split Categorized by Clusters

Cluster	Trip Pair Number									
	1	2	3	4	5	6	7	8	9	10
	Edm/Cal	Edm/GrP	Edm/FMM	Edm/FMK	Cal/GrP	Cal/Let	Cal/FMM	Cal/FMK	Cal/Red	Cal/Lld
Day 1										
Air	**7%**	**2%**	**15%**	**25%**	**20%**	**2%**	**41%**	**34%**	**2%**	**17%**
Ground (no stops)	78%	74%	68%	60%	80%	66%	59%	66%	75%	33%
Ground (with stops)	15%	23%	18%	15%	0%	32%	0%	0%	24%	50%
Ground (Total)	**93%**	**98%**	**85%**	**75%**	**80%**	**98%**	**59%**	**66%**	**98%**	**83%**
Day 2										
Air	**29%**	**21%**	**47%**	**47%**	**74%**	**7%**	**79%**	**81%**	**11%**	**n/a**
Ground (no stops)	54%	64%	35%	40%	26%	67%	20%	19%	68%	**n/a**
Ground (with stops)	14%	14%	14%	14%	0%	24%	2%	0%	18%	**n/a**
Ground (long 2)	3%	0%	4%	0%	0%	2%	0%	0%	4%	**n/a**
Ground (Total)	**71%**	**79%**	**53%**	**53%**	**26%**	**93%**	**21%**	**19%**	**89%**	**n/a**

As discussed in Section 3.1.2, the raw anonymous cellular data was checked between Edmonton and Calgary using field tests. The field tests confirmed the accuracy of the location of a cell phone user to the records (by tower location) in the raw data. The rest of the province was not checked using field tests; however, since the data was demonstrated to be of high fidelity in the Edmonton Calgary corridor, it is reasonable to assume that the data is of good quality for the rest of the province as well. As discussed in Section 1.1, accuracy is less of an issue when looking on trips at a larger scale (i.e. city to city) versus within a city.

There are several interesting observations that can be made from this analysis. First, the mode share for air trips all increased from Day 1 to Day 2. This is a similar trend to the mode split found in Section 4.1, and once again, could be because more people are choosing to fly instead of drive due to the winter season on Day 2. Certain trip pairs appear to have an uncharacteristically high proportion of trips made by air on Day 2. However, this is possibly explained by the fact that Day

2 is a winter day, and it is more likely that the trip pairs of greatest distance such as #5. Calgary – Grande Prairie, #7. Calgary – Fort McMurray, and #8. Calgary – Fort MacKay occurred by air. On the other hand, Figure 11 showed that the distinction between air and ground clusters was not as clear as Day 1, so it is also possible that some of the inferred air trips are in fact drive trips. The limited data makes it unclear whether this is a seasonal variation in travel patterns, or data sampling variations or errors.

Fort McMurray and Fort McKay have the special characteristic that many people work there on a fly-in-fly-out shift schedule, with flights that transport workers in and out of the area (King, 2014). We can see that the trip pairs between Edmonton – Fort McMurray, Edmonton – Fort McKay, Calgary – Fort McMurray, Calgary – Fort McKay have relatively high mode shares by air, which approximately doubles on Day 2. The air mode share from Calgary is also higher than from Edmonton, most likely due to Calgary being further away and less workers choosing to commute by drive on their days off.

For Day 1, there is a 98% drive mode split for trip pairs: #2. Edmonton – Grande Prairie, #6. Calgary – Lethbridge, and #9. Calgary – Red Deer. With the exception of Trip Pair #2, Calgary/Lethbridge and Calgary/Red Deer both have a high ratio of average flight time over average ground travel time (0.4 and 0.5 respectively). First, this may indicate that Lethbridge and Red Deer is so close to Calgary that most trips between them are made by ground. Second, the high ratio of the flight time over the ground travel time may make it harder to cluster properly. Mode splits between different cities will vary due to distance and air service characteristics. For example, Calgary to Red Deer is an 80-minute drive or a 40-minute flight. However, the pre-boarding time needed for taking a flight adds an hour and half to the travel time; flying may end up being the mode of travel with the highest time cost compared with driving.

In conclusion the mode split results showed interesting results that were not all expected. It wasn't expected that the mode split would vary greatly between the two days; a more robust dataset could be used in the future to identify whether the changes are due to seasonal variations. The mode split results show favourable results in that trends can be explained based on the season and specific trip pair's distances apart.

5. Conclusions

An overview of the research conducted in this thesis and key findings are presented in this chapter. The contributions of this research and suggestions for future work are also provided.

5.1 Research Overview

This research investigates the use of anonymous cellular data for intercity travel patterns in the province of Alberta. Intercity travel is a relatively understudied area in comparison to urban transportation systems. There has traditionally been less focus placed on intercity travel because it accounts for a much smaller portion of total travel. Rising intercity demands have increased the impacts of congestion and emissions from intercity travel, and researchers and governmental agencies are recognizing the need to better understand intercity travel. However, traditional methods of data collection such as for OD flows through household surveys or roadside surveys is time consuming and expensive and public agencies have historically prioritized survey data collection within their jurisdiction, typically urban boundaries. This thesis looks at applying passive data, specifically anonymous cellular data to identify intercity travel patterns. This is done by extracting trip volumes and inferring the mode split of whether an intercity trip was conducted by ground or air. Intercity trips were first extracted between Edmonton and Calgary. Two methods were utilized to infer the trip mode, first by categorizing the travel times using upper and lower limits, second by hierarchical clustering of the travel times. This analysis was then expanded to all intercity trips between cities in Alberta. Hierarchical clustering of a rescaled travel time was conducted for all intercity trip pairs that had direct air service between them.

5.2 Research Findings

Intercity trips were first extracted between Edmonton and Calgary. The extracted trips show that the majority of trips are of duration between 2 – 3 hrs, as well as a smaller portion between 0 – 1 hrs. This provides some confirmation to the validity of anonymous cellular data to identify intercity travel trends, as a direct drive trip between the two cities take approximately 3 hours, and a flight takes 45 minutes from takeoff to landing. Intercity trips between all cities in Alberta (along with urban service area Fort McMurray and oil sands camps Fort MacKay) were then extracted using a similar methodology. OD tables of these extracted trips were shown in percentage. There were ten city pairs that have direct air service between them. It was found that the majority of intercity trips occur between Edmonton – Calgary, and Fort McMurray – Fort MacKay, followed by trips from Calgary – Red Deer, Calgary – Lethbridge, and Edmonton – Red Deer. Overall, the data shows that larger cities have more trips, and the distance between cities also affected the share of trips (i.e. smaller cities had the most trips to and from cities nearby, sparsely located cities had very few trips anywhere).

For extracted trips between Edmonton and Calgary, after inferring the trip mode by two methods, it was found that hierarchical clustering is a more appropriate method. Hierarchical clustering of trips less than eight hours show that there are distinct clusters that represent air trips, ground trips, and longer ground trips (possibly with stops made). Mode split was ranged from 12 – 25% air and 88 – 75% ground for the two days from hierarchical clustering.

All extracted trips from these ten cities were clustered together to infer their trip mode. Since different city pairs have different travel times, the travel times were rescaled based on the average ground travel time so that they could be clustered together. The mode share for air trips were higher for all trip pairs on Day 2 than Day 1, which suggest that the winter season increases mode share by air. Certain trip pairs had much higher share of air trips, which was observed mostly in trip pairs that were furthest apart.

5.3 Research Contribution

My work in this thesis provides contribution to researchers studying long-distance travel, passive data applications in transportation, as well as government transportation agencies and other industry practitioners. First it adds to the limited literature on identifying OD flows for long-distance travel. As well, estimating OD flows between urban areas have not been conducted on a larger scale, this research provides a simple methodology to extract trips over a provincial network of cities. The methodologies are less complex than what has been conducted among other studies that identify OD flows, but arguably this work shows how it is possible to gain a picture of movement in Alberta with a larger geographical scope and limited data set than what other studies have attempted thus far.

This thesis also contributes to the application of using anonymous cellular data for mode split. It builds upon H. Wang et al.'s (2010) work that shows that trip mode can be inferred with just the observed travel time through clustering techniques. They inferred trip modes on an intraurban scale using CDRs, whereas my work is on an intercity scale with more detailed data. My work infers trip mode between air and ground, which has not been done before.

Currently there is no province-wide travel survey in Alberta, so my work provides some new information about intercity travel characteristics in Alberta, and methods that may be applied to larger more comprehensive cellular data sets than the sample tested as part of my research. This is valuable to governmental agencies such as Alberta Transportation (AT), who manage the highway network across Alberta as OD flows between urban zones in Alberta would be new knowledge. The mode split analysis would be useful for airport authorities that would like to know the share of people that choose to fly over drive between certain cities. As well, this information can be valuable to AT given recent events such as the 2016 Fort McMurray wildfire. I was part of the research team that studied the evacuation and re-entry and anonymous cellular data would have provided useful information such as how evacuees (that stayed in Alberta) spread out across the province, and where evacuees were re-entering from.

Overall, though there are limitations to the dataset, my work demonstrates the potential with anonymous cellular data to provide valuable information to intercity travel. It demonstrates that from anonymous cellular data, useful information about long-distance travel such as intercity OD flows and air/ground mode splits can be obtained.

5.4 Research Limitations and Future Work

Perhaps the biggest limitation in this work is the limited dataset. Two individual days of sample data is not enough to make significant and conclusive inferences about intercity travel in the province. Secondly, the data obtained contained only a sample of all records from the cellular service provider, therefore representing a small sample size. However, as discussed, even though the dataset was limited, this research demonstrated the value in extracting useful intercity travel information.

Anonymous cellular data is a passive data source, which does not provide any information about the user's characteristics, including ones that we know are critical for explaining travel behavior on both urban and intercity scales. As well, passive data comes with uncertainty in the sampling variances, errors, and noise. Thus it is more challenging to process and extract meaningful information from this data source.

In the future, it is recommended that a more robust, comprehensive dataset be used in the analysis. This means a dataset with longer continuous collection periods and not just a sample of the cellular service provider's data. Longer periods of data can help identify to what extent sampling biases affects the data, as well as whether there are any seasonal trends in travel. As well, home location zones (for example on a TAZ level) can then be identified, which can be linked to socio-demographic information from household travel surveys. This could then be used for all forms of travel modelling such as intercity mode choice models, intercity demand models, airport leakage models, and urban models.

Future work also includes integrating vehicle count data with the anonymous cellular data to expand the extracted intercity trips to actual population volumes. As well, more exploration in the data could be conducted to improve the accuracy of the results by better identifying what is data noise and errors. There would also be value in providing more documentation of anonymous cellular data and the methodology with practitioners as a target audience.

As discussed in Section 2.3, there are more and more opportunities ranging from using apps and services on our smartphones to the Internet to social media platforms from which we could glean information about travel behavior (including long-distance travel). There are now ways to collect passive data sources without their current limitations, such as through the smartphone apps that can record trips passively and prompt the user for trip details after. I believe that the use of anonymous cellular data will continue to change and there will be more ways to integrate this data with other sources to obtain data that has even more value, with all the benefits of being faster and less expensive than traditional household travel surveys to collect.

REFERENCES

Aamaas, B., Borken-Kleefeld, J., & Peters, G. P. (2013). The climate impact of travel behavior: A German case study with illustrative mitigation options. *Environmental Science and Policy, 33*, 273-282.

Abdelwahab, W. M. (1991). Transferability of intercity disaggregate mode choice models in Canada. *Canadian Journal of Civil Engineers, 18*, 20-26.

Aggarwal, C. C., & Reddy, C. K. (2014). *Data Clustering Algorithms and Applications.* Taylor and Francis Group.

Alberta Government. (2017). *Population*. Retrieved June 24, 2017, from Economic Dashboard: http://economicdashboard.alberta.ca/Population

Alexander, L., Jiang, S., Murga, M., & Gonzalez, M. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies, 58*, 240-250.

Anderson, M., & Simkins, J. (2012). *Development of Long Distance Multimodal Passenger Travel Modal Choice Model.*

Ashiabor, S., Baik, H., & Trani, A. (2007). Logit Models for Forecasting Nationwide Intercity Travel Demand in the United States. *Transportation Research Record: Journal of the Transportation Research Board*, 1-12.

Badger, E. (2013, September 26). *Dazzling Timelapse Videos of Millions of FourSquare Check-Ins.* Retrieved August 25, 2017, from Citylab: https://www.citylab.com/life/2013/09/dazzling-timelapse-videos-millions-foursquare-check-ins/7034/

Bar-Gera, H. (2007). Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C*, 380-391.

Bekhor, S., Cohen, Y., & Solomon, C. (2013). Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions. *Journal of Advanced Transportation, 47*(4), 435-446.

Bhat, C. R. (1995). A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research Part B, 29*(6), 471-483.

Bradley, M., & Daly, A. (1997). Estimation of logit choice models using mixed stated preference and revealed preference information. *Understanding travel behaviour in an era of change*, 209-232.

Bureau of Transportation Statistics. [BTS] (2003). *Highlights of the 2001 National Household Travel Survey.* Washington, D.C.: US Department of Transportation.

Caceres, N., Romero, L. M., & Benitez, F. G. (2013). Inferring origin-destination trip matrices from aggregate volumes on groups of links: a case study using volumes inferred from mobile phone data. *Journal of Advanced Transportation*, 650-666.

Caceres, N., Romero, L., & Benitez, F. d. (2012). Traffic flow estimation models using cellular phone data. *IEEE Transactions on Intelligent Transportation Systems, 13*(3), 1430-1441.

Caceres, N., Wideberg, J., & Benitez, F. (2008). Review of Traffic Data Estimations Extracted from Cellular Networks. *IET Intelligent Transport Systems, 2*(3), 179-192.

Caceres, N., Wideberg, J., & Benitez, F. G. (2007). Deriving origin-destination data from a mobile phone network. *IET Intelligent Transport Systems, 1*(1), 15-26.

Calabrese, F., & Ratti, C. (2006). Real Time Rome. *Networks and Communication Studies, 20*(3-4), 247-258.

Calabrese, F., Di Lorenzo, G., L., & Ratti, C. (2011). Estimating Origin Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing, 99*.

Canadian Radio-television and Telecommunications Commission. (2014). *Communications Monitoring Report 2014: Canadians at the centre of the communications system.* Government of Canada. Retrieved June 25, 2016, from http://www.crtc.gc.ca/eng/publications/reports/PolicyMonitoring/2014/cmr2.htm

CBC News. (2016). *Calgary airport records a record-breaking 2015.* Retrieved July 14, 2016, from CBC News: http://www.cbc.ca/news/canada/calgary/calgary-airport-record-year-2015-1.3421478

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software, 61*(6).

Cho, H. D. (2013). *The Factors that Affect Long-Distance Travel Mode Choice Decisions and Their Implications for Transportation Policy (PhD thesis).* University of Florida.

City of Edmonton. (2017). *City Sector Profiles.* Retrieved June 25, 2016, from City of Edmonton: https://www.edmonton.ca/business_economy/demographics_profiles/city-sector-profiles.aspx

Colak, S., Alexander, L., Alyim, B., Mehndiratta, S., & Gonzalez, M. (2015). Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation Research Record: Journal of the Transportation Research Board*(2526), 126-135.

Commercial Aviation Safety Team. (2013, April). *Phase of Flight Definitions and Usage Notes.* Retrieved June 26, 2017, from National Transportation Safety Board: https://www.ntsb.gov/investigations/data/documents/datafiles/PhaseofFlightDefinitions.pdf

Dargay, J. M., & Clark, S. (2012). The determinants of long distance travel in Great Britain. *Transportation Research Part A*(46), 576-587.

Edmonton International Airport. (2016). *Passenger Statistics*. Retrieved July 14, 2016, from EIA: http://corporate.flyeia.com/business-at-the-airport/air-service-development/passenger-statistics

Election and Census Services. (2017). *2016 Municipal Census Results*. Retrieved April 9, 2017, from City of Edmonton: https://www.edmonton.ca/city_government/facts_figures/municipal-census-results.aspx

Elliot, A., & Woodward, W. A. (2007). *Statistical Analysis Quick Reference Guidebook with SPSS Examples.* Sage Publications, Inc.

Gonzalez, M., Hidalgo, C., & Barabasi, A. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779-782.

Government of Alberta. (2016, December 22). *Types of Municipalities in Alberta*. Retrieved March 22, 2017, from Alberta Municipal Affairs: http://municipalaffairs.gov.ab.ca/am_types_of_municipalities_in_alberta

Government of Alberta. (2017, March 17). *Alberta Municipal Affairs profiles summary report: cities.* Retrieved March 22, 2017, from Alberta Government Open Goverment: http://municipalaffairs.alberta.ca/cfml/MunicipalProfiles/basicReport/CITY.PDF

Grayson, A. (1981). Disaggregate Model of Mode Choice in Intercity Travel. *Transportation Research Record 385*, 36-42.

Hodson, H. (2016, July 20). *Baidu uses millions of users' location data to make predictions*. Retrieved August 25, 2017, from New Scientist: https://www.newscientist.com/article/2098206-baidu-uses-millions-of-users-location-data-to-make-predictions/

Holz-Rau, C., Scheiner, J., & Sicks, K. (2014). Travel distances in daily travel and long-distance travel: what role is played by urban form? *Environment and Planning A, 46*, 488-507.

Horak, R. (2006). *Telecommunications and Data Communications Handbook*. Retrieved July 5, 2016, from https://books.google.ca/books?id=dO2wCCB7w9sC&pg=PA111&dq=%22Call+detail+record%22&hl=en&redir_esc=y#v=onepage&q=%22Call%20detail%20record%22&f=false

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Estimating human trajectories and hotspots through mobile phone data. *Computer Networks, 64*, 296-307.

Hui, K., Wang, C., Kim, A., & Qiu, T. (2017). Investigating the use of anonymous cellular phone data to determine intercity travel volumes and modes. *Transportation Research Board 2017 Annual Meeting Compendium of Papers.* Washington, D.C.

Iqbal, M., Choudhury, C., Wang, P., & Gonzalez, M. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies, 40*, 63-74.

Itsubo, S., & Hato, E. (2006). Effectiveness of Household Travel Survey Using GPS-Equipped Cell Phones and Web Diary: Comparative Study with Paper-Based Travel Survey. *Transportation Research Board 85th Annual Meeting.* Washington, D.C.: Transportation Research Board.

Johnson, S. C. (1967). Hierarchical Clustering Schemes. *Psychometrika, 32*(3), 241-254.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding Human Mobility from Twitter. *pLos one, 10*(7).

Kaski, S. (1997, March 31). *Clustering methods*. Retrieved April 25, 2017, from http://users.ics.aalto.fi/sami/thesis/node9.html

King, T. (2014, November 6). *Fly-In, Fly-Out is our biggest threat*. Retrieved July 18, 2017, from Fort McMurray Today: http://www.fortmcmurraytoday.com/2014/11/06/fly-in-fly-out-is-our-biggest-threat

Koppelman, F. S. (1989). Multidimensional Model System for Intercity Travel Choice Behavior. *Transportation Research Record 1241*, 1-8.

Koppelman, F. S., & Wen, C.-H. (2000). The paired combinatorial logit model: properties, estimation and application. *Transportation Research Part B, 34*, 75-89.

Kuhnimhof, T., Collet, R., Armoogum, J., & Madre, J.-L. (2009). Generating Internationally Comparable Figures on Long-Distance Travel for Europe. *Transportation Research Record: Journal of the Transportation Research Board*(2105), 18-27.

Kunzmann, M., & Daigler, V. (2013). *2010-2012 California Household Travel Survey Final Report.* California Department of Transportation.

LaMondia, J. J., Aultman-Hall, L., & Greene, E. (2014). Long-Distance Work and Leisure Travel Frequencies. Ordered Probit Analysis Across Non-Distance-Based Definitions. *Transportation Research Record: Journal of the Transportation Research Board*, 1-12.

LaMondia, J. J., Moore, M., & Aultman-Hall, L. (2015). Modeling Intertrip Time Intervals Between Individuals' Overnight Long-Distance Trips. *Transportation Research Record*(2495), 23-31.

Lee, R. J., Sener, I. N., & Mullins III, J. A. (2016). An Evaluation of Emerging Data Collection Technologies for Travel Demand Modeling: From Research to Practice. *TRB 2016: 95th Annual Meeting of the Transportation Research Board.* Washington, D.C.

Liu, H. X. [Henry X.], Danczyk, A., Brewer, R., & Starr, R. (2008). Evaluation of Cell Phone Traffic Data in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board*(2066), 1-7.

Liu, H. [Haobing], Xu, Y., Stockwell, N., Rodgers, M. O., & Guensler, R. (2016). A comparative life-cycle energy and emissions analysis for intercity passenger transportation in the U.S. by aviation, intercity bus, and automobile. *Transportation Resarch Part D, 48*, 267-283.

McGuckin, N., & Nakamoto, Y. (2004). Trips, Chains and Tours - Using and Operational Definition. *National Household Travel Survey Conference.*

Miller, E. (2004). The trouble with intercity travel demand models. *Transportation Research Record: Journal of the Transportation Research Board*, 94-101.

Morrison, S., & Winston, C. (1985). An Econometric Analysis of the Demand for Intercity Passenger Transportation. *Research in Transportation Economics, 2*, 213-237.

Noulas, A., Mascolo, C., & Frias-Martinez, E. (2013). Exploiting foursquare and cellular data to infer user activity in urban environments. *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on. 1*, pp. 167-176. IEEE.

Oum, T. H., & Gillen, D. W. (1983). The Structure of Intercity Travel Demands in Canada: Theory Tests and Empirical Results. *Transportation Research B, 17B*, 175-191.

Poon, L. (2017, January 11). *Finally, Uber Releases Data to Help Cities with Transit Planning*. Retrieved August 25, 2017, from Citylab: https://www.citylab.com/transportation/2017/01/finally-uber-releases-data-to-help-cities-with-transit-planning/512720/

Qiu, Z., Jin, J., Cheng, P., & Ran, B. (2007). State of the Art and Practice: Cellular Probe Technology Applied in Advanced Traveler Information Systems. *TRB 86th Annual Meeting Compendium of Papers.* Washington, D.C>: Transportation Research Board of the National Academies.

Ridout, R., & Miller, E. J. (1989). A disaggregate logit model of intercity common carrier passenger modal choice. *Canadian Journal of Civil Engineering, 16*(4), 568-575.

Ritter, C., & Greene, E. (2017). Have Smartphone, Will Travel: Long Distance Travel Surveys with Smartphone GPS. *TRB 2017 Annual Meeting*. Washington, D.C.: Transportation Research Board.

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology, 17*(4), 688-690.

S A, R., Karim, M. A., Qiu, T. Z., & Amy, K. (2015). Origin-Destination Trip Estimation from Anonymous Cell Phone and Foursquare Data. *Transportation Research Board 94th Annual Meeting.* Washington, D.C.: Transportation Research Board.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web* (pp. 851-860). Raleigh: ACM.

Schiffer, R. G. (2012). *NCHRP Report 735: Long-Distance and Rural Travel Transferable Parameters for Statewide Travel Forecasting Models.* Washington, D.C.: National Academies Press.

Schlaich, J., Otterstatter, T., & Friedrich, M. (2010). Generating trajectories from mobile phone data. *Proceedings of the 89th annual meeting compendium of papers.* Transportation Research Board of the National Academies.

Sharp, J., Jonaki, B., Giesbrecht, L., Memmott, J., Khan, M., & Roberto, E. (2004). A Picture of Long-distance Travel Behavior of Americans Through Analysis of the 2001 National Household Travel Survey. *National Household Travel Survey Conference.*

Skaltsas, G. (2011). *Analysis of Airline Schedule Padding on U.S. Domestic Routes. Master's Thesis.* Massachusetts Institute of Technology.

Sohn, K., & Kim, D. (2008). Dynamic Origin-Destination Flow Estimation Using Cellular Communication System. *IEEE Transactions on Vehicular Technology, 57*(5), 2703-2713.

Sohn, T., Varshavsky, A., LaMarca, A. C., Choudhury, T., Smith, I., Consolvo, S., . . . de Lara, E. (2006). Mobility Detection Using Everyday GSM Traces. *Lecture Notes in Computer Science*, (pp. 212-224).

Stantec. (2007). *Provincial Highway Service Classification Final Report.* Alberta Transportation.

Statistics Canada. (2007, January 24). *Communications for the Travel Survey of Residents of Canada*. Retrieved from Statistics Canada: http://www23.statcan.gc.ca/imdb-bmdi/document/3810_D3_T9_V1-eng.pdf

Statistics Canada. (2011, February 4). *Population, urban and rural, by province and territory*. Retrieved June 24, 2017, from Statistics Canada: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo62j-eng.htm

Statistics Canada. (2015, November 27). *Census metropolitan area (CMA) and census agglomeration (CA)*. Retrieved May 21, 2017, from Statistics Canada: http://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo009-eng.cfm

Statistics Canada. (2016). *Travel Survey of Residents of Canada - 2016*. Retrieved April 9, 2017, from Statistics Canada: http://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&a=1&&lang=en&Item_Id=296747

Statistics Canada. (2017a, January 23). *Census Profile, 2016 Census*. Retrieved April 24, 2017, from Statistics Canada: http://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=POPC&Code1=0292&Geo2=PR&Code2=47&Data=Count&SearchText=Fort%20McMurray&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=0292&TABID=1&wbdisable=true

Statistics Canada. (2017b, March 9). *Population of census metropolitan areas*. Retrieved June 26, 2017, from Statistics Canada: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo05a-eng.htm

Statistics Canada. (2017c, March 1). *Travel Survey of Residents of Canada (TSRC)*. Retrieved April 13, 2017, from Statistics Canada: http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3810

Statistics Canada. (n.d.). *Table 1-1 - Air Carrier Traffic at Canadian Airports*. Retrieved July 27, 2017, from Statistics Canada: http://www.statcan.gc.ca/pub/51-203-x/2015000/t002-eng.htm

Steenbruggen, J., Tranos, E., & Nijkamp, P. (2015). Data from mobile phone operators: A tool for smarter cities? *Telecommunications Policy, 39*, 335-346.

Stopher, P., & Collins, A. (2005). Conducting a GPS prompted recall survey over the internet. *Transportation Research Board 84th Annual Meeting.* Washington, D.C.: Transportation Research Board.

Stopher, P., & Prashker, J. (1976). Intercity Passenger Forecasting: The Use of Current Travel Forecasting Procedures. *Annual Meeting of the Transportation Research Forum*, (pp. 67-75).

TEMS, Inc. and Oliver Wyman. (2008). *Market Assessment of High Speed Rail Service in the Calgary-Edmonton Corridor.*

The Canadian Press. (2017, February 8). *Census 2016: Population of metropolitan Calgary outpaced Canada's growth rate*. Retrieved April 9, 2017, from Global News: http://globalnews.ca/news/3235534/census-2016-population-of-metropolitan-calgary-outpaced-canadas-growth-rate/

Train, K. (2009). *Discrete Choice Methods with Simulation.* Cambridge University Press.

U.S. National Park Service. (2017). *World Physical Map.* esri. Retrieved April 26, 2017, from https://services.arcgisonline.com/ArcGIS/rest/services/World_Physical_Map/MapServer

United States Department of Transportation [USDOT]. (n.d.). *Long-Distance Travel.* Retrieved April 7, 2017, from Bureau of Transportation Statistics: https://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/highlights_of_the_2001_national_household_travel_survey/html/section_03.html

Wang, H., Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. *2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, (pp. 318-323).

Wang, M.-H., Schrock, S. D., Broek, N. V., & Mulinazzi, T. (2013). Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data. *International Journal of Intelligent Transportation Systems Research, 11*(2), 76-86.

Wang, R. Q., & Taylor, J. E. (2014). Quantifying Human Mobility Perturbation and Resilience in Hurricane Sandy. *PLoS one, 9*(11).

Wen, C.-H., & Koppelman, F. S. (2001). The generalized nested logit model. *Transportation Research Part B, 35*, 627-641.

Wilson, F. R., Damodaran, S., & Innes, J. D. (1990). Disaggregate mode choice models for intercity passenger travel in Canada. *Canadian Journal of Civil Engineering*, 184-191.

Yang, F. [Fan], Jin, P. J., Wan, X., & Li, R. (2013). Dynamic Origin-Destination Travel Demand Estimation using Location Based Social Networking Data. *92nd Transportation Research Board Annual Meeting.* Washington, D.C.: Transportation Research Board.

Yang, F. [Fei], Yao, Z. J., & Yang, D. (2016). Performance Evaluation of Handoff-Based Cellular Traffic Monitoring Systems Using Combined Wireless and Traffic Simulation Platform. *Journal of Intelligent Transportation Systems*, 1547-2450.

Ygnace, J. (2001). *Travel time/speed estimates on the French Rhone corridor network using cellular phones as probes.* Lyon: SERTI V Program. System for Traffic Information and Positioning (STRIP) Project.

Zhang, L., Southworth, F., Xiong, C., & Sonnenberg, A. (2012). Methodological Options and Data Sources for the Development of Long-Distance Passenger Travel Demand Models: A Comprehensive Review. *Transport Reviews, 32*(4), 399-433.

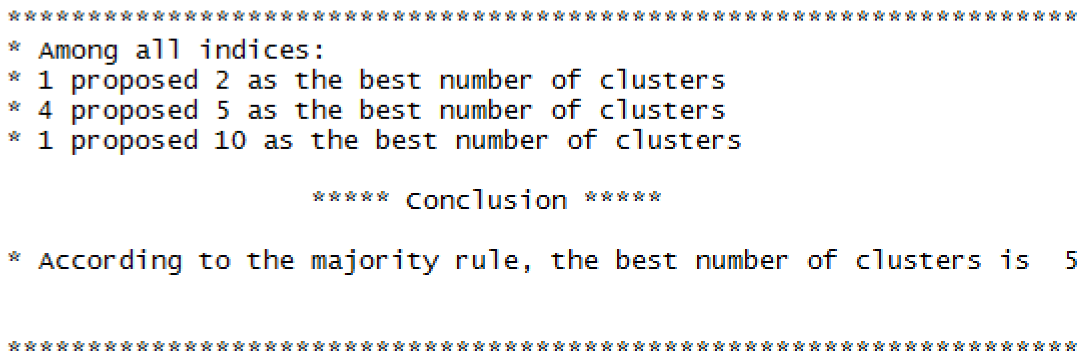
Zhang, Y., Qin, X., Dong, S., & Ran, B. (2010). Daily O-D Matrix Estimation using Cellular Probe Data. *89th Annual Meeting Transportation Research Board.* Washington, D.C.

APPENDIX A: OPTIMAL NUMBER OF CLUSTERS FROM R

The NbClust package in R (https://cran.r-project.org/web/packages/NbClust/NbClust.pdf) was used to check whether the number of clusters chosen in the mode split analysis was optimal for the data. NbClust provides 30 indices to determine the optimal number of clusters (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). A minimum to maximum number of clusters is set, as well as the clustering technique (ex. average linkage in hierarchical clustering). 26 indices are applied, and a summary of the number of indices that suggest k number of clusters is provided. The results did not always show that three clusters is optimal, however it is ultimately up to the analyst to decide what are the best results and for our analysis, three clusters were chosen.

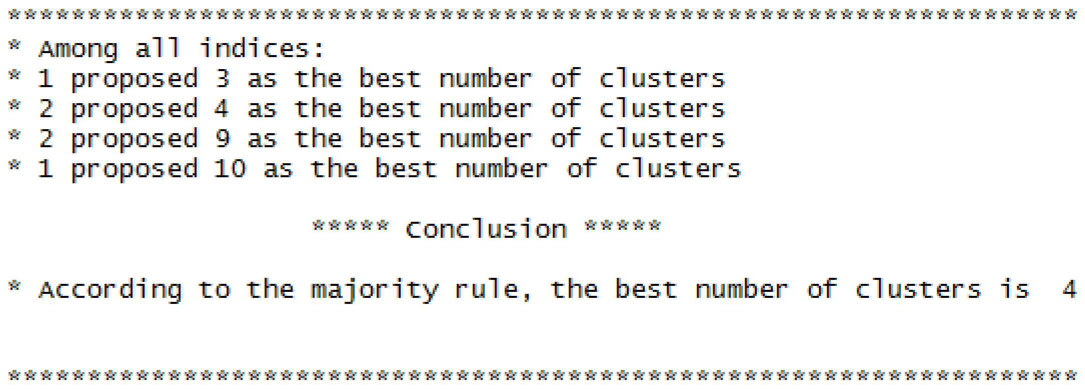
R Code:

```
if(!require(devtools)) install.packages("devtools")

devtools::install_github("kassambara/factoextra")

pkgs <- c("cluster",  "NbClust")

install.packages(pkgs)

library(readxl)

library(factoextra)

library(NbClust)

library(ggplot2)

library(cluster)

# Day 1 Trips between Edmonton - Calgary

day1EDMCAL<- read_excel("C:/Users/huikat/OneDrive -
ualberta.ca/Research/Thesis/Data/1 - CP Data/day1ClustersEDmCAL.xlsx", +
sheet = "Sheet1")

trips <- day1EDMCAL$TT_hr

result <- NbClust(trips, distance = "euclidean", min.nc=2, max.nc=10, method
= "average", index = "all")
```



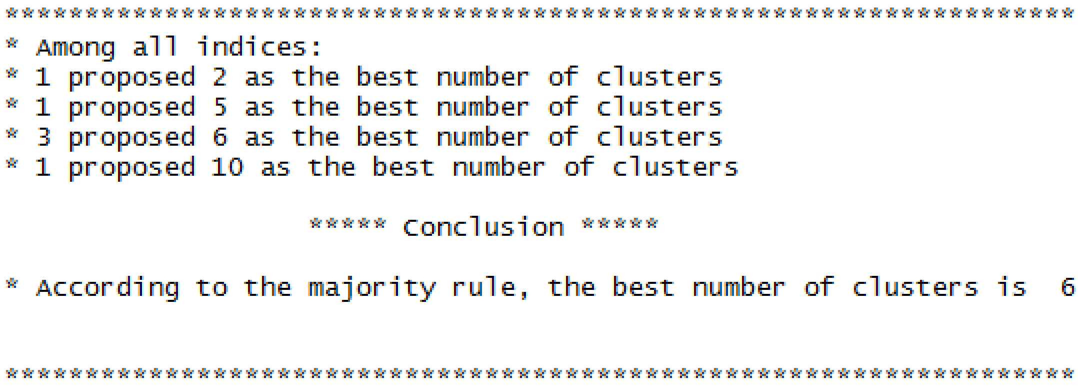
```
*******************************************************************
* Among all indices:
* 1 proposed 2 as the best number of clusters
* 4 proposed 5 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  5


*******************************************************************
```

```
# Day 2 Trips between Edmonton - Calgary

day2EDMCAL<- read_excel("C:/Users/huikat/OneDrive -
ualberta.ca/Research/Thesis/Data/1 - CP Data/day2ClustersEDMCAL.xlsx", +
sheet = "Sheet1")

trips <- day2EDMCAL$TT_hr

result <- NbClust(trips, distance = "euclidean", min.nc=2, max.nc=10, method
= "average", index = "all")
```



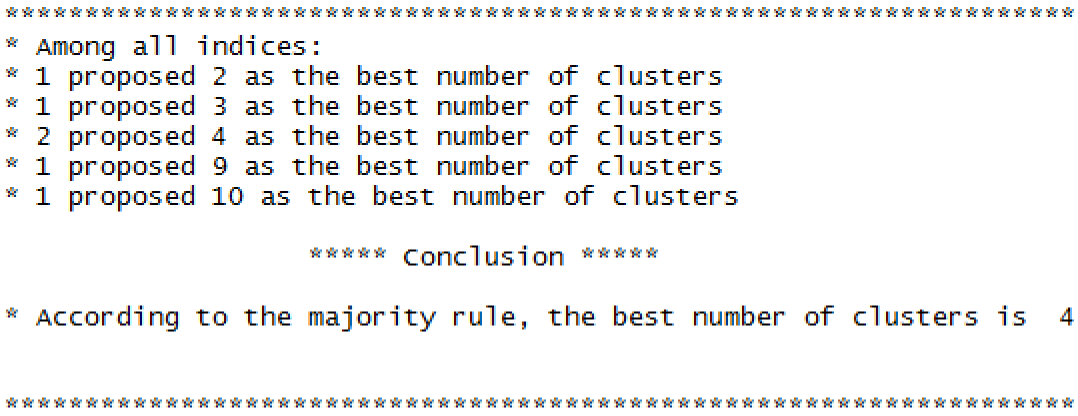
```
*******************************************************************
* Among all indices:
* 1 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 2 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4


*******************************************************************
```

```
# Day 1 Trips of 10 intercity trip pairs

day1AB <- read_excel("C:/Users/huikat/OneDrive -
ualberta.ca/Research/Thesis/Data/3 - CP Data/clustersDay1.xlsx", + sheet =
"Sheet1")

trips <- day1AB$TT_NormGr

result <- NbClust(trips, distance = "euclidean", min.nc=2, max.nc=10, method
= "average", index = "all")
```



```
*******************************************************************
* Among all indices:
* 1 proposed 2 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 3 proposed 6 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  6


*******************************************************************
```

```
# Day 2 Trips of 10 intercity trip pairs

day2AB <- read_excel("C:/Users/huikat/OneDrive -
ualberta.ca/Research/Thesis/Data/3 - CP Data/clustersDay2.xlsx", + sheet =
"Sheet1")

trips <- day2AB$TT_NormGr

result <- NbClust(trips, distance = "euclidean", min.nc=2, max.nc=10, method
= "complete", index = "all")
```



```
*******************************************************************
* Among all indices:
* 1 proposed 2 as the best number of clusters
* 1 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  4


*******************************************************************
```

APPENDIX B: CLUSTERING WITH RESCALED TRAVEL TIMES BY AVERAGE FLIGHT TIMES

Rescaling the travel times with the average flight time per trip pair is conducted without favourable results. The following formula is used:

$$Travel\ Time_a^i = \frac{Actual\ Travel\ Time_x^i}{Average\ Flight\ Time_x}$$

Where

The superscript $i = 1 \ldots n$ denotes the trip observation number.

The subscript a denotes the travel time normalized by the average flight time.

The subscript x indicates the trip pair number that the trip observation belongs to.

Hypothetically, we should expect that a cluster of air trips would have a mean of 1, and a cluster of drive trips would have a mean somewhere in the range of 2 – 5. Using the filtered data set on the condition of $Travel\ Time_g^i$ less than three, the clustering process needed five clusters to observe a cluster that may represent fly trips. The possible fly trip cluster also contains trips that look more like they belong to the drive cluster (circled in red in Figure B12).

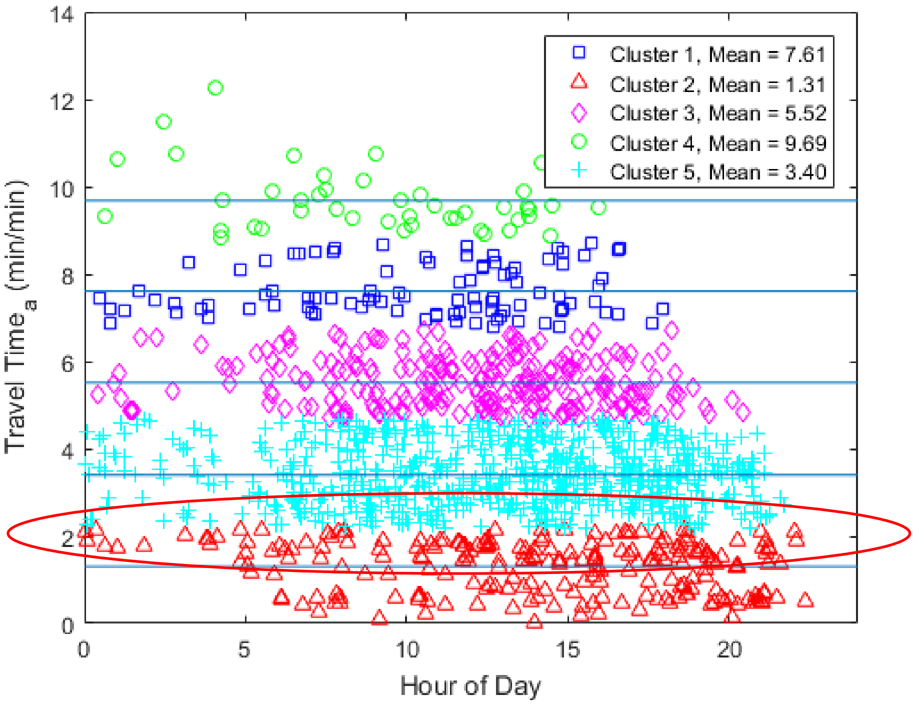


Figure B12 Five clusters from Travel Time$_a$ on Day 1

Thus, another filter is done on these trips to focus on trips with a smaller Travel Time$_a$ value to see whether a distinct cluster for air trips can be observed. This time, a filter of $Travel\,Time_a^n$ less than six is applied, and a first attempt with two clusters is shown in Figure B13 below.

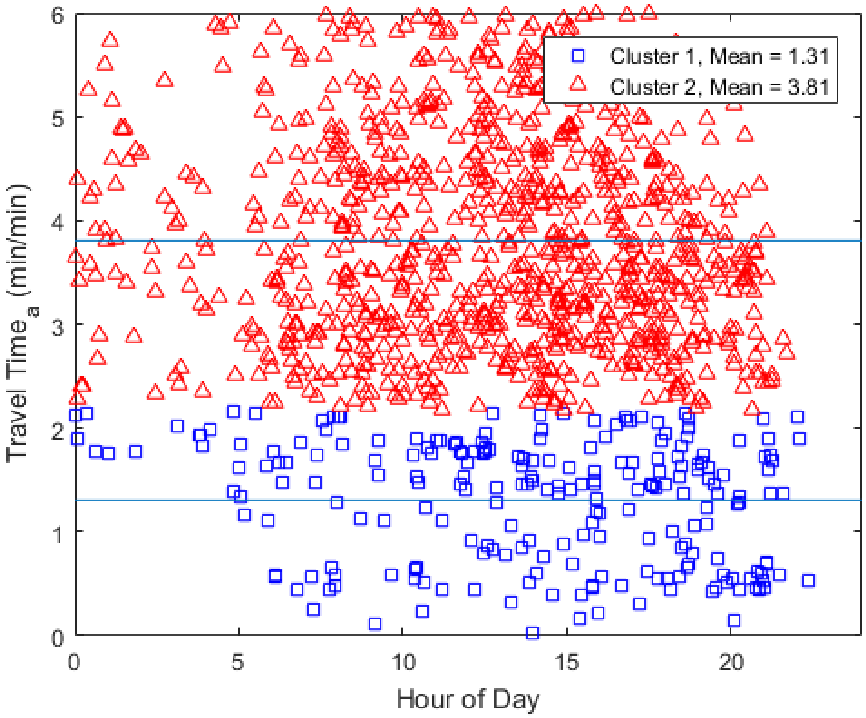


Figure B13 Two clusters with Travel Timea < 6 on Day 1

From this, we identify two clusters that look like something we would expect, with the first cluster having a mean of 1.31 and a second cluster at 3.81. Overall, it appears that the rescaled travel time based on average ground travel times is clusters better than the average flight times. A possible reason for this is because average ground travel times are more stable, especially between many of the cities where traffic volume is not high. Thus the travel time rescaled by average flight time was not used in the analysis.