10-5-2017

Studying Regional and Cross Border Freight Movement Activities with Truck GPS Big Data

Kevin Gingerich

University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation

Gingerich, Kevin, "Studying Regional and Cross Border Freight Movement Activities with Truck GPS Big Data" (2017). Electronic Theses and Dissertations. 7262.
https://scholar.uwindsor.ca/etd/7262

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000 ext. 3208.
Studying Regional and Cross Border Freight Movement Activities with Truck GPS Big Data

By

Kevin Gingerich

A Dissertation
Submitted to the Faculty of Graduate Studies through the Department of Civil and Environmental Engineering in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the University of Windsor

Windsor, Ontario, Canada

2017

© 2017 Kevin Gingerich
Studying Regional and Cross Border Freight Movement Activities
with Truck GPS Big Data

by

Kevin Gingerich

APPROVED BY:

__________________________________________________
M. Roorda, External Examiner
University of Toronto

__________________________________________________
P. Graniero
Department of Earth and Environmental Sciences

__________________________________________________
C. Lee
Department of Civil and Environmental Engineering

__________________________________________________
W. Anderson
Department of Civil and Environmental Engineering

__________________________________________________
H. Maoh
Department of Civil and Environmental Engineering

August 28, 2017
DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION

I. Co-Authorship

I hereby declare that this thesis incorporates material that is result of joint research, as follows:

Chapter 2 of the thesis utilized material from a report submitted to Transport Canada and a published article in the journal *Transportation Research Part C* that were both co-authored by Dr. Hanna Maoh and Dr. William Anderson. Chapter 3 and Chapter 5 of the thesis are derived from separate published articles in *Transportation Research Record* that were also co-authored with Hanna Maoh and William Anderson. Chapter 4 and Chapter 6 of the thesis are based on draft papers that have not yet been submitted for publication, but are co-authored by Hanna Maoh.

In each of the above cases, the primary data analysis and writing was performed by the author. The co-authors, namely Dr. Hanna Maoh, provided frequent supervisory input on the overall direction of the papers and provided editing support for the text and revisions before final acceptance for publication. In addition, Dr. William Anderson participated in the discussion of the ideas and the overall direction of the research questions.

I am aware of the University of Windsor Senate Policy on Authorship and I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.
II. Previous Publication

This thesis includes three original papers that have been previously published/submitted for publication in peer reviewed journals, as follows:

<table>
<thead>
<tr>
<th>Thesis Chapter</th>
<th>Publication title/full citation</th>
<th>Publication status</th>
</tr>
</thead>
</table>

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.
III. General

I declare that, to the best of my knowledge, my thesis does not infringe upon anyone’s copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.
ABSTRACT

This dissertation utilizes an existing GPS data source to create and analyze a dataset of processed truck trips. The original data was generated for the purpose of fleet management by GPS transponders installed on Canadian owned trucks. These vehicles provide a critical service by fulfilling the economic need to move goods from one location to another. This thesis subsequently re-purposes the GPS pings as a form of opportunistic data to enrich the current state of knowledge regarding freight movement patterns.

The first sections of this thesis are dedicated towards understanding the GPS data and devising processing methods needed to convert raw data into a suitable dataset of truck trips. Due to the nature of the topic, a geographic perspective was integral to this work to properly mine the data for useful information. For example, a new application of entropy based on the variety and distribution of carriers stopping at a location was created to assist with the classification of stop events. The data processing resulted in an approximate sample size of 245,000 trips per month from September 2012 to December 2014 and the month of March 2016. The volume of data and level of detail provides information that has not been available to date, which includes trip origins and destinations, associated industry, observed routes, and border crossing time/location if the trip was international.

The processed trips derived from GPS data are applied towards a better understanding of inter-regional and cross-border truck movements. This area is under-represented due to the difficulties in obtaining long-haul trip data where trucks move through multiple jurisdictions. These difficulties are compounded for international trips
since the study area spans multiple nations. The processed truck trips are utilized to identify the spatial patterns of truck movements at specific border crossings between Canada and the U.S. including the Ambassador Bridge, Blue Water Bridge, and Peace Bridge. The choice of border crossing is also investigated using a specific case study of trucks travelling between Toronto, Ontario, and Chicago, Illinois. Finally, the observed trips from origin to destination allows for an analysis of delays at single locations (the border crossing) as well as their impact on the total trip.

These applications represent a small part of the full potential that passive GPS data can provide after sufficient processing is applied. It is the hope of this author that these efforts can contribute towards the state of practice in transportation as GPS data becomes increasingly available to researchers. The work presented in this thesis illustrates how such GPS data can be used as a viable source to fill in gaps in knowledge. While traditional data collection techniques will remain a necessary facet of transportation research in the foreseeable future, information generated passively by users every day provides a new source of data that is characteristically large (in terms of volume and spatio-temporal coverage) and cost-effective.
ACKNOWLEDGEMENTS

While the list of all those who have helped me in one way or another would be too long for this section, I would like to acknowledge the contributions of certain individuals. First and foremost, I would like to thank my supervisor, Hanna Maoh, for his persistent support and mentorship during my tenure as a graduate student. His efforts continually improved on my work and pushed me further than I would have ever expected. I feel rather privileged to have had him as my supervisor/mentor for all of these years. I would also like to extend acknowledgements to the other committee members for this thesis. Chris Lee and Bill Anderson have both been involved in many activities related to my thesis. Bill provided constant support throughout my graduate studies and graciously accepted me into the Cross-Border Institute (CBI) fold where this PhD research began and found its purpose. Chris was notably the first teacher I had in the transportation field, which has eventually led to the point where I am now. I would also like to thank Phil Graniero for his help on this thesis committee and Matthew Roorda for serving as the external committee member.

I would also like to acknowledge the support of Transport Canada which provided some initial funding and temporarily loaned the GPS data to CBI. In particular, this work was made possible by the efforts of Louis-Paul Tardif and Andrew Carter. I would also like to thank the financial contributions provided by NSERC through a PGS-D scholarship and also FedDev Ontario.

Numerous other individuals have also contributed in some way to this thesis. I would like to acknowledge the many contributions provided by Shakil Khan, who I have had the pleasure of sharing an office with for many years. I would also like to
acknowledge the varied contributions of some staff at the Cross-Border Institute including Goldie, Patricia, and Maureen. I would also like to thank Haibin Dong for the software development of the firm identification tool. Students that have helped in some regard include Rahaf, Kenneth, Flavio, and Zachary. Finally, I would like to thank my parents (I still have that key) and siblings for their constant love and support.
TABLE OF CONTENTS

DECLARATION OF CO-AUTHORSHIP / PREVIOUS PUBLICATION .................... iii
ABSTRACT ........................................................................................................ vi
ACKNOWLEDGEMENTS ............................................................................ viii
LIST OF TABLES ........................................................................................... xvi
LIST OF FIGURES ......................................................................................... xvii

CHAPTER 1 INTRODUCTION ......................................................................... 1
  1.1 Preface ................................................................................................. 1
  1.2 Transportation in Canada ..................................................................... 2
  1.3 Freight Transportation Modelling ....................................................... 6
  1.4 Traditional Freight Modelling Data Sources ....................................... 8
      1.4.1 Firms and Employment ................................................................. 8
      1.4.2 Freight Vehicle Trips .................................................................... 9
  1.5 Opportunistic Big Data ......................................................................... 10
      1.5.1 Application of GPS data in Transportation Research ................ 12
      1.5.2 Advantages of Opportunistic GPS data ........................................ 15
      1.5.3 GPS Data Availability and Trends ............................................... 16
  1.6 Primary Data Source ........................................................................... 17
  1.7 Thesis Research Questions and Objectives ......................................... 18
      1.7.1 Mining truck GPS Big Data .......................................................... 18
      1.7.2 Apply the GPS Dataset to Better Understand Cross-border and Inter-
                    regional Truck Movements ......................................................... 22
  1.8 Thesis Outline ..................................................................................... 24
2.1 Introduction .................................................................................................................. 31

2.2 GPS Data Characteristics .............................................................................................. 33
  2.2.1 GPS Data Attributes .................................................................................................. 33
  2.2.2 GPS Data Accuracy .................................................................................................... 34

2.3 Identifying Carrier Shipping Depots .............................................................................. 37
  2.3.1 Identifying Shipping Depots Using Clustering ............................................................ 38
  2.3.2 Identifying Shipping Depots Using Microsoft SQL Server ........................................ 39
  2.3.3 Shipping Depot Identification Results ....................................................................... 41
  2.3.4 Validations of Results ............................................................................................... 42

2.4 Stop Events .................................................................................................................... 43

2.5 Classifying Stop Events ................................................................................................ 46
  2.5.1 Definitions for Stop Event Classification .................................................................... 46
  2.5.2 Previous Approaches to Stop Event Classification .................................................... 47
  2.5.3 Entropy Method for Classifying Stop Events ........................................................... 50
  2.5.4 Entropy Validation Results ....................................................................................... 54
  2.5.5 Entropy Method Discussion ..................................................................................... 59
  2.5.6 Additional Classification Methods ............................................................................ 62

2.6 Stop Event Industry Identification ................................................................................ 64
  2.6.1 Firm Identification Tool ............................................................................................ 66
  2.6.2 Firm Datasets Tested .................................................................................................. 67

2.7 Data Processing Approach .............................................................................................. 69
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>Chapter 2 References ..................................................................</td>
<td>72</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction .............................................................................</td>
<td>74</td>
</tr>
<tr>
<td>3.2</td>
<td>Background ...............................................................................</td>
<td>76</td>
</tr>
<tr>
<td>3.3</td>
<td>Methods of Analysis ..................................................................</td>
<td>79</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Location and Type of Stops ..................................................</td>
<td>79</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Identifying Valid Trips .......................................................</td>
<td>81</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Border Crossing Time ................................................................</td>
<td>85</td>
</tr>
<tr>
<td>3.4</td>
<td>Discussion and Results ................................................................</td>
<td>86</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Spatial Distribution of Trips ...............................................</td>
<td>88</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Density Maps of International Trip Productions ..........................</td>
<td>92</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Border Crossing Time ..................................................................</td>
<td>96</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Effect of Distance on Crossing Times ......................................</td>
<td>99</td>
</tr>
<tr>
<td>3.5</td>
<td>Conclusions ..............................................................................</td>
<td>100</td>
</tr>
<tr>
<td>3.6</td>
<td>Chapter 3 References ..................................................................</td>
<td>102</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction .............................................................................</td>
<td>104</td>
</tr>
<tr>
<td>4.2</td>
<td>Background ...............................................................................</td>
<td>107</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Freight Trip Generation ..........................................................</td>
<td>107</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Sample Expansion ......................................................................</td>
<td>109</td>
</tr>
<tr>
<td>4.3</td>
<td>Primary Data and Biases ................................................................</td>
<td>110</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Distance Bias ...........................................................................</td>
<td>111</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Industry Bias ..........................................................................</td>
<td>112</td>
</tr>
</tbody>
</table>
7.3.2 Representation of Industry and Goods..........................................................188
7.3.3 Truck Tours and Multi-modal Analysis.........................................................190
7.3.4 Data Integration...............................................................................................191
7.3.5 Route Choice Modelling................................................................................192
7.4 Final Remarks ....................................................................................................193
7.5 Chapter 7 References .......................................................................................194
APPENDIX A ............................................................................................................196
VITA AUCTORIS ......................................................................................................232
LIST OF TABLES

Table 1-1: Sample information accessible from raw GPS data .............................................. 18
Table 2-1: Lateral GPS errors ........................................................................................................ 37
Table 2-2: Validated industries from 1,000 randomly selected stop events ...................... 65
Table 3-1: Summary of data and trips by month (2013) ............................................................... 87
Table 3-2: Crossing times by aggregate destination industry ..................................................... 98
Table 4-1: Average trip rates by jurisdiction .............................................................................. 115
Table 4-2: Initial trip rates for Canadian firms by industry ......................................................... 116
Table 5-1: Variables used in the crossing choice model .............................................................. 136
Table 5-2: Discrete choice model results .................................................................................... 141
Table 5-3: Sensitivity of Ambassador crossing times ................................................................. 143
Table 6-1: Border crossing statistics (in minutes) for July, 2013 ................................. 162
LIST OF FIGURES

Figure 1-1: Canadian GDP by year ................................................................. 4
Figure 1-2: Proportion of Canadian international trade by mode ..................... 4
Figure 1-3: Canada-U.S. land border crossings .............................................. 5
Figure 1-4: Sample truck tour ..................................................................... 7
Figure 1-5: Increase in GPS based transportation articles over time .............. 12
Figure 1-6: Connecting pings with a straight line (left) or map-matched route (right) . 15
Figure 2-1: Original purpose of GPS data ................................................... 33
Figure 2-2: GPS ping dispersion at the Ambassador Bridge ........................... 35
Figure 2-3: GPS ping dispersion along a Highway 401 corridor .................... 36
Figure 2-4: Example of GPS points (left) and corresponding density surface (right) .... 39
Figure 2-5: Primary carrier truck yards by Canadian census division ............... 41
Figure 2-6: Distance based dwell time calculation ........................................ 44
Figure 2-7: Entropy index (EI) interpretations .............................................. 52
Figure 2-8: Maximum entropy values by number of truck carriers ................ 53
Figure 2-9: Entropy values for two carriers with trucks stopped at a given location .... 54
Figure 2-10: Histogram of entropy results for all clusters and shipping depots ...... 56
Figure 2-11: Stacked bar chart of stop purposes by value of entropy ................. 57
Figure 2-12: Stop purpose for entropy ranging from 2.5 to 3 (150 point sample) ...... 58
Figure 2-13: Stop purpose for full entropy range (250 point sample) ............... 58
Figure 2-14: GPS industry classification tool .............................................. 67
Figure 2-15: Microsoft SQL scripts for GPS processing ............................... 69
Figure 3-1: Principle attributes processed for each GPS derived trip ............... 76
Figure 3-2: Zoning system applied to GPS based trips................................. 80
Figure 3-3: Example of initial potential trips from Zone A to Zone B................. 82
Figure 3-4: Example of a final trip from Zone A to Zone B............................. 82
Figure 3-5: Example of a multi-zone trip .................................................. 83
Figure 3-6: Relationship between effective vehicle speed and distance............... 84
Figure 3-7: Geofence configurations for the Ambassador and Blue Water bridges..... 86
Figure 3-8: Origin trips using the Ambassador Bridge by census division (2013) ....... 89
Figure 3-9: Origin trips using the Blue Water Bridge by census division (2013) .... 90
Figure 3-10: Origin trips in US zones through the Ambassador Bridge (2013)......... 91
Figure 3-11: Origin trips in US zones through the Blue Water Bridge (2013)......... 91
Figure 3-12: Canadian density of truck trips crossing the Ambassador Bridge ......... 93
Figure 3-13: U.S. density of truck trips crossing the Ambassador Bridge ............... 93
Figure 3-14: Canadian density of truck trips crossing the Blue Water Bridge ......... 94
Figure 3-15: U.S. density of truck trips crossing the Blue Water Bridge ............... 94
Figure 3-16: Canadian density of truck trips crossing the Peace Bridge ............... 95
Figure 3-17: U.S. density of truck trips crossing the Peace Bridge..................... 95
Figure 3-18: Border crossing time distributions (2013) .................................. 96
Figure 3-19: Average crossing time at the two bridges classified by trip distance .... 100
Figure 4-1: Distribution of trips grouped by distance ..................................... 112
Figure 4-2: GPS sample representation of firms by industry ........................... 113
Figure 4-3: Flow chart outlining trip expansion ............................................ 114
Figure 4-4: Total trip count and trip rate relationship for Ontario census divisions.... 117
Figure 4-5: IPF method inputs and output .................................................... 118
Figure 4-6: Shortest path routes between Ontario zones ........................................ 120
Figure 4-7: Correspondence between OD trips and survey stations ......................... 121
Figure 4-8: Scatterplot of observed (GPS) and expected (CVS) trips .......................... 123
Figure 4-9: Validation results by survey station ...................................................... 123
Figure 5-1: Northern and southern routes between Toronto and Chicago ................ 129
Figure 5-2: Average times at the Ambassador (top) and Blue Water (bottom) .......... 137
Figure 5-3: Crossing volumes towards Canada (top) and the U.S. (bottom) .......... 138
Figure 5-4: Distribution of monthly crossings by day ............................................. 140
Figure 6-1: Border geofences for three crossing locations ..................................... 159
Figure 6-2: Origin frequency of trips crossing three bridges (2013) ......................... 160
Figure 6-3: Border delay as a proportion of the trip travel time for Canada bound
(top) and U.S. bound (bottom) trips ................................................................. 165
Figure 6-4: Delay as a proportion of full trips ....................................................... 168
Figure 6-5: Average trip delay proportion by zone ................................................. 169
Figure 6-6: Zones with the highest average trip delays ......................................... 170
Figure 7-1: Road link volumes from one week of GPS trips .................................. 193
CHAPTER 1
INTRODUCTION

1.1 Preface

The increasing pace of technological development, adoption, and connectivity is providing remarkable opportunities for new discoveries in various disciplines. The advantage brought by the widespread of information and communication technologies is a massive volume of data that was absent from the academic scene but started emerging only about a decade ago. The consensus is that this “Big Data” is reshaping the way human and natural systems are studied.

As in the case of most disciplines, the usage of Big Data in transportation research is non-trivial and forms a major challenge for researchers and practitioners alike. Digging into the Big Data frontier draws from a number of areas besides civil engineering such as spatial science and computer science. Despite the challenge, the success in mining information from these opportunistic Big Data is expected to give rise to one of the major innovations in the field of freight transportation as the data provides an expanded understanding on the nature of transportation activities.

One such data source that is gaining increased attention in transportation research is Global Positioning System (GPS) technology, which has been broadly adopted for navigation and vehicle tracking. This thesis presents research focused on utilizing an existing Big Data source derived from GPS devices that track the movements of approximately 60,000 Canadian trucks. The objective of the conducted research is to
understand the nature of long distance inter-regional trips across Canada and between Canada and the U.S. (i.e. cross-border truck movements). A fundamental departure from previous work is the large geographical and temporal scope that the utilized Big Data provides. The latter characteristics allow timely answers to pressing research questions that could not have been addressed otherwise. As such, the conducted research in this thesis provides several new discoveries within the realm of freight transportation.

The remainder of the introduction chapter is organized into seven sections (excluding references) that provide context to this research and outline the research questions and objectives that are addressed in the thesis.

1.2 Transportation in Canada

The transportation field is primarily concerned with the movement of people and goods from one location to another over space. These movements are created by activities that occur over the course of any given day such as work, shopping, and recreation (in the case of personal travel), or the manufacturing of materials for other processed goods or final consumption (in the case of freight travel). The generated activities subsequently create demand for some method of travel to facilitate the flow of people or goods. In order to satisfy this demand, government agencies (or occasionally private firms) provide infrastructure including roadways (and sidewalks/trails), rail, ports (air and sea), and pipelines.

The modern transportation field came into prominence in the middle of the twentieth century as larger populations and urbanization began increasing the demand for interconnected transportation infrastructure. In the U.S., the interstate highway system was conceived in 1956 with the Federal Aid Highway Act (Weingroff, 1996). In Canada,
the Trans-Canada Highway Act was signed in 1949 to create a highway connecting all of Canada’s provinces that officially opened in 1962 and was completed in 1971 (MacLeod, 2014). The primary purpose of transportation engineering during this period was viewed as the enhancement of traffic mobility (Levinson, 2003), leading to increasing investments in the capacity of highways.

Today, the role of transportation engineering is more profound as it is required to balance traffic mobility with sustainable objectives including economic efficiency, environmental sensitivity, and social responsibilities (Litman, 2016). But major roads and highways still persist as the primary transportation infrastructure supply for the movement of people and goods. The historical trend of highway investment has contributed towards trucking as a predominant mode of transport that accounts for 31% of Canada’s commercial transportation sector by gross domestic product (GDP). The remaining modes by air, rail, and marine based travel represent 12%, 11%, and 2% of Canada’s commercial transportation GDP respectively (Transport Canada, 2012). By volume, 72% of domestic goods are transported by trucks, while rail and marine modes only haul 21% and 7%, respectively (Transport Canada, 2015).

Within the greater economy, the transportation and warehousing sector accounts for a substantial 4.3% of Canada’s GDP directly and 10% indirectly (Transport Canada, 2015). While the Canadian economy has experienced recent volatility (due to the financial crisis in 2008/2009 and the large drop of oil prices in 2014), Figure 1-1 demonstrates the rapid economic growth that has taken place since 2002. A rising GDP places a strong emphasis on transportation to move an increasing number of goods (Maoh et al., 2016).
As part of the Great Lakes corridor shared by Canada and the U.S. (Sands, 2009), the Province of Ontario is highly dependent on trade with the U.S. as the largest destination for its exports (representing 80.88% of all exported goods), followed by the United Kingdom (6.68%), Mexico (1.68%), and China (1.37%) (Government of Ontario, 2017). This trade relationship makes Ontario an extremely trade oriented entity with both exports and imports individually representing approximately 31% of the Ontario GDP (Anderson, 2012). Moreover, trade between Canada and the U.S. relies heavily on trucks as the dominant method of transport, representing 56.5% of all modes of travel in 2011 as shown in Figure 1-2 (Transport Canada, 2011). Therefore domestic and Canada-U.S. trade are both highly dependent on trucks as a major source of commercial transportation to ship goods.
While trucking is the predominant mode of transport for Canada-U.S. trade, the geography of the Great Lakes imposes limited possibilities for truck border gateways to travel between Ontario and the U.S. As a result, a very high proportion of trucks between Canada and the U.S. flow through a small number of border crossings. This includes the Ambassador Bridge, Peace Bridge, and Blue Water Bridge representing 28%, 17%, and 13% of Canada-U.S. truck volumes (Maoh et al., 2016), respectively, as shown in Figure 1-3. These crossing locations consequently represent extremely critical links for the Canadian economy.

Figure 1-3: Canada-U.S. land border crossings

Source: Maoh et al. (2016)
1.3 Freight Transportation Modelling

The most commonly used modelling approach in transportation is the 4 step urban transportation planning system (UTPS). This system encompasses (1) trip generation, (2) trip distribution, (3) modal split, and (4) network assignment (Ortuzar and Willumsen, 2011). The building block of the UTPS method is an aggregate (zonal) approach to simulating the flow of traffic, with correspondingly modest data requirements.

For freight transportation, an alternative aggregate approach to the first two/three phases of the UTPS is the generation of goods movement demand from an economic perspective such as Multi-Regional Input-Output (MRIO) models (Maoh et al., 2008; Bachmann et al., 2014) or Computable General Equilibrium (CGE) models (Roberts et al., 2014). The economic approach reflects the production and transportation of goods to fulfill the requirements of other goods production (intermediate demand) and consumption for households, private investment, and exports (final demand) (Miller and Blair, 2009). However, this approach still typically simplifies to aggregate patterns exhibited by commercial vehicles.

As freight modelling has gained increased attention over the past 15 years, state of the art modelling practices (particularly in academic circles) have shifted towards activity based (microscopic) freight models that represent the individual movements of commercial vehicles and goods. This approach is considered superior to aggregate models for both passenger and commercial vehicle modelling due to the ability to track the true behaviour and variability of individual agents and their evolution over time (Miller et al., 2004). The range of microscopic models for freight movements predominantly belong to one of two classes (Chow et al., 2010): logistics models used to
track the movement of goods across a supply chain (Raothanachonkun et al., 2008; Holguín-Veras and Thorson, 2003) and truck tour models (Hunt and Stefan, 2007).

Supply chain models focus on the movement of goods between various agents involved in the supply chain process and the interactions between these agents. For example, the interactions between shippers, carriers (Cavalcante and Roorda, 2013, Liedtke, 2009) and customers (Liedtke, 2009) have been modelled in detail.

In contrast to the supply chain models that focus on the entire journey encountered by goods, truck tour models study the movement patterns of individual trucks. A truck tour is typically defined as a round trip where a truck leaves a starting point to perform one or more stops before returning to the initial establishment in a process known as trip chaining. Figure 1-4 shows a sample scenario of a truck performing a multi-leg tour employing trip chaining. An example of truck tour models can be found in Hunt and Stefan (2012).

![Figure 1-4: Sample truck tour](image)

*Source: Gingerich and Maoh (2015)*
Across all three advanced freight modelling approaches described above, the origin and destination (OD) of trips is a particularly important type of information. The knowledge gained from inter-regional OD freight data also provides immense information to policy makers and planners by identifying the current patterns of trade connected to various jurisdictions.

1.4 Traditional Freight Modelling Data Sources

While activity based approaches (such as the supply chain and truck tour models) are able to capture detailed transportation patterns of individual vehicles and goods, they also rely on detailed data that may be unavailable and/or too costly to collect. Two major types of data are integral to modelling freight movements. These are discussed below in Sections 1.4.1 and 1.4.2.

1.4.1 Firms and Employment

Firms and employment by industry are important statistical controls that have been used extensively in freight models since they are the agents that produce/refine the goods that are eventually transported. Aggregate employment data (and occasionally restricted access to individual data) can be obtained from periodic population census information (Maoh and Kanaroglou, 2009). Moreover, aggregate information on industries can also be obtained from economic account information (Bachmann et al. 2015). Business pattern data can often be obtained at some scale of area and industry (Oliveira-Neto et al. 2012) to provide some aggregate control of employment for freight models. Information on the location of individual firms can be obtained from commercial organizations such as Google or InfoCanada (Ferguson et al., 2012). The actual location
of firms provides very useful data for activity based freight transportation research. As an example, the InfoCanada dataset is an important part of this thesis for (i) the determination of industry associated with stop events of trucks and subsequent truck trips, (ii) the identification of rest stops, and (iii) the calculation of trip rates per firm derived from the GPS data.

In addition to firm data, their relationships between each other can also require data inputs, particularly for supply chain models modelling the interactions between various agents. This information may be difficult to obtain from revealed preference data based on observable results. A common alternative source of data in this case (and other transportation issues) comes from stated preference surveys, which pose theoretical questions to respondents. For example, the Freight Market Interactions Simulation (FREMIS) supply chain model described by Cavalcante and Roorda (2013) utilized a stated preference survey to obtain information describing the freight market behaviour of shippers and carriers.

1.4.2 Freight Vehicle Trips

Similar to firm and employment data, information on freight trips can also be obtained from a variety of sources. Aggregate vehicle count information can often be obtained from municipal or provincial level entities, such as the major highway counts in Ontario collected by the Ministry of Transportation Ontario (MTO). Moreover, waybill information can be used to identify details of shipments. For example, Brown and Anderson (2015) obtained waybill information from the Statistics Canada Trucking Commodity Origin-Destination Survey to study the cost of trading across the Canada-US border. Other survey information may also be available including commercial vehicle
surveys of carriers, receivers (Holguin-Veras et al. 2011), truck drivers, and retailers (Nuzzolo and Comi, 2014). Such surveys in the past have typically been derived from in-person, telephone, or online questionnaires. Moreover, trip diary surveys can be utilized to capture the individual characteristics of vehicle routes. These interviews could be based on past behavior based on memory recall (Wang and Hu, 2012), current activities based on active data logging (Leore, 2015), or theoretical routes if stated preference surveys are conducted (Toledo et al., 2013). The active data logging is utilized by a Canadian Vehicle Use study where a GPS datalogger is mailed to survey participants and installed on the vehicle for 21 days before mailing back (Leore, 2015). While surveys are certainly useful for generating new data, the costs and participation rates associated with them can be prohibitive, leading to potentially small sample sizes. For example, the Canadian Vehicle Use survey had captured a limited duration of movements for 2,000 heavy trucks since inception in 2011 to reporting in January 2015 (Leore, 2015).

1.5 Opportunistic Big Data

An alternative to creating new data from surveys is the adaptation of data already created for other purposes. Such information can be defined as opportunistic data (International Transport Forum, 2015). This data recycling provides a relatively inexpensive approach for the researcher since the owner generates a smaller amount of extra revenue for data that has already served its original purpose. The primary drawback to utilizing opportunistic data is that in most cases, the obtained records require extensive data mining and processing to create a dataset that matches the needs of the researcher. However, emerging efforts (such as those found in this thesis) contribute towards overcoming this limitation.
Another aspect of opportunistic datasets is that they are often characterized as Big Data. For example, the majority of the population in Canada and the U.S. possess cell phones that allow for portable communications but also create data records that may contain the location of users. Transportation planners can utilize this data by purchasing it from companies such as AirSage, who collect 15 billion data points per day in the U.S. that are derived from triangulating cell phone locations using cellular towers (FHWA, 2014). A common definition of Big Data introduced by Doug Laney from the META group includes 3 V’s: volume, velocity, and variety (Laney, 2001). Other additions to this definition have later been added including veracity and value (USDOT, 2014). Big Data will likely become an increasingly important tool in the near future to help understand and solve difficult transportation problems:

“The combination of low-cost and widespread sensing (much of it involving personal devices), the steep drop in data storage costs and the availability of new data processing algorithms improves our ability to capture and analyze more detailed representations of reality. Today these representations augment traditional sources of transport data collection. In the future they will likely replace them.” - International Transport Forum (2015).

An example of opportunistic Big Data in the context of truck movements is Global Positioning System (GPS) pings. Such data is often generated from transponders installed on carrier trucks to perform real-time management of their fleet and potentially optimize routes. As a data source that observes the movement of vehicles, the generated data can also be utilized by other interested parties. As an example of the growing popularity of GPS data for transportation, a search of journal articles was conducted using Engineering Village (2017) for journal articles that had both ‘transportation’ and ‘GPS’ in the title, subject, or abstract. The overall number has grown from 10 articles
published in 1990 to 343 articles published in 2015 as shown in Figure 1-5. While the number of articles with just ‘transportation’ as the keyword has also increased (from 9,078 published in 1990 to 53,752 published in 2015), the overall proportion of GPS papers increased from 0.11% to 0.64%, confirming the growing trend of GPS data as a tool for transportation research.

![Figure 1-5: Increase in GPS based transportation articles over time](image)

*Source: Adapted from Engineering Village (2017)*

1.5.1 Application of GPS data in Transportation Research

Studies utilizing GPS data for freight transportation can be categorized into several major groups depending on the purpose of their application. These groups include truck tour models, analysis of reliability along freight corridors, and trip routing. These applications are discussed in the paragraphs below.

While early truck tour models utilized traditional survey information for model calibration (such as Hunt and Stefan, 2007), GPS data can be utilized as an alternative source of data. Greaves and Figliozzi (2008) applied 1 week of GPS data pertaining to 30 trucks to create truck tours in Melbourne, Australia. The study presented information on the truck tours such as trip distances, stops per tour, tour distance, etc. Of particular importance to this thesis, the study noted a potential application of the passive GPS data
to generate detailed origin-destination matrices. Kuppam et al. (2014) created a microscopic truck tour model for Phoenix, Arizona. With respect to model approach, it presents similarities to Hunt and Stefan (2007) but employed GPS data as the primary source of information instead of traditional surveys. Ma et al. (2016) utilized GPS data collected over the course of 3 days to analyze the trip chaining behaviour of trucks and classify these movement patterns into four categories. The latter was performed using PAM clustering based on differences occurring in average trip chains per truck, average stops per trip chain (tour), average dwell time, average trip chain length and average trip length. Sharman (2014) utilized three months of GPS data to create two truck modelling components for a larger modelling framework of trips and tours: activity duration models and inter-arrival duration models.

GPS data has not only been used for direct model calibration, but has become a beneficial addition to measures of reliability. A survey conducted by the Transportation Association of Canada found that provinces have a strong interest in measures of transportation performance that included reliability and mobility, but there is no consistent measure used across Canada (TAC, 2006). GPS data is particularly well suited for reliability measures calculated across a very wide area, allowing for potential unification across multiple jurisdictions that does not currently exist.

Wang et al. (2015) measured reliability of trucks using GPS data along a 3.5 mile section of Interstate 5 in Seattle by presenting an improved method of spot-speed detection and comparing with other reliability metrics including the coefficient of variation, buffer time index, and truck reliability index. Anderson and Coates (2010) measured crossing times at the four major crossings between Southern Ontario and the
U.S. utilizing GPS data over a one year period. The study focuses on the Buffer Index as a measure of reliability and the implications of such a measure on the planned arrival time for supply chains. While few studies have utilized GPS data for border crossings thus far, other examples include studies by Goodchild et al. (2008) and McCord et al. (2010). However, none of these studies combined information gathered at the border with the rest of the activities performed by the observed vehicles, which becomes an important part of this thesis. The extent of research opportunities with the GPS data depends on its quality and characteristics including the size of the dataset, accuracy of the location results, and frequency of the GPS pings. The properties of the GPS data used in this thesis are discussed in more detail in Section 1.6 and in Chapter 2.

Revealed preference (observed) route choice studies are scarce in transportation literature due to traditional data collection difficulties (Prato, 2009). While GPS data exists originally as a set of pings (described more in Section 1.6 and Chapter 2), the points can be converted into routes utilizing map-matching (Dalumpines, 2014, Dhakar, 2012). Map-matching is a process where point events (in this case derived from GPS pings generated by a moving vehicle) are converted into a line event that is matched to a digital representation of the road network and identifies the route of the vehicle. As evidenced by Figure 1-6, the actual route taken by a vehicle is inherently more complicated than a straight line connecting the two pings (points). As a result, the GPS data is an alternative to traditional surveys that provides detailed information on routing choices made by vehicles. Li et al. (2005) utilized GPS data to study the route choice of commuters during the morning period. The study included 10 days of data for 182 drivers. Quattrone and Vitetta (2011) utilized both roadside surveys and GPS data to
provide information on route choices of trucks for 52 monitored routes on the Italian road network.

![Connection of pings with a straight line or map-matched route](image)

**Figure 1-6: Connecting pings with a straight line (left) or map-matched route (right)**

1.5.2 Advantages of Opportunistic GPS data

The increasing popularity of GPS data for research can be attributed to its particular advantages and growing availability. The advantages of opportunistic GPS data include:

- **Large study area** – since the data is passively generated by vehicles (as dynamic probes), long distance trips will correspondingly result in data that covers large distances / areas.

- **Large sample size** – GPS data has become an integral part of most vehicles on the road today. As a result, large samples are becoming increasingly possible.

- **Cost effective** – a massive amount of information can be gathered from data that already exists in the transportation industry and does not require the substantial costs of establishing surveys.

- **Track changes over time** – The GPS data represents a passive source that is continuously created over time. Subject to acquisition availability, the data can be used to observe trends across months or longer.
• **Observes entire trips** – The ability of GPS data to observe entire trips can provide immense amounts of information. Many facets of individual trips can be selected from a GPS dataset of trips to suit various research/project needs.

1.5.3 GPS Data Availability and Trends

The application of GPS data as an opportunistic source for freight transportation modelling requires a willingness from companies that own such data to provide (or sell) this information. In the U.S., a popular source of truck based GPS data can be obtained from the American Transportation Research Institute (ATRI). Examples of studies that use ATRI data include Flaskou et al. (2015), Kuppam et al. (2014), and Bernardin et al. (2015). In Canada, several studies have utilized data provided by XRS (previously known as Turnpike Global Technologies) including Anderson and Coates (2010) and Sharman (2014). With respect to data size, Sharman (2014) utilized XRS GPS data pertaining to 77 carriers and 1,618 trucks for a study area encompassing the Greater Toronto Area.

Shaw Tracking has more recently made their GPS data available to government agencies such as Transport Canada and the Ministry of Transportation Ontario (MTO). The data obtained from Shaw Tracking contains over 850 trucking carriers in total, representing over 60,000 Canadian trucks that move across North America. To the best of my knowledge, the Shaw Tracking data is by far the largest source of GPS Big Data available in Canada. The released records consist of approximately 1.1 billion records for any given year and provides the primary source of data presented in this thesis. Notably, the Shaw GPS data after extensive processing provides valuable information on Canadian trucks that move within Canada and travel into the U.S.
1.6 Primary Data Source

The GPS data presented in this thesis pertains to data collected from Shaw Tracking over several periods of time. The initial raw GPS data occurred over a time period inclusive of September 2012 to December 2014. This GPS data was temporarily loaned to researchers at the Cross-Border Institute (CBI) at the University of Windsor as part of a collaboration with Transport Canada (TC). As per the arrangement, the raw data from this collaboration has been deleted from CBI storage and is no longer used for further analysis. However, due to the success of the initial projects using GPS data, CBI has purchased permanent access to more recent GPS data for the time period covering July, 2015 to March, 2016. In some cases, only a subset of data based on a given time period has been used for specific applications in the thesis to reduce the overall processing burden. As such, the time period of data will be declared in the dissertation where appropriate. Overall, the data tends to be consistent over time. For example, aggregate counts of trips crossing individual road segments were compared between the first two weeks in March 2016. The correlation of trip volumes between the two weeks for any road links with at least one trip was calculated to be 99.21%.

For all time periods, the raw data was received as a comma delimited text file. Due to the size of data, files were typically received separately for each month. For example, the March, 2016 text file is 4.88 GB in size. In the raw data file, each row (i.e. GPS record) corresponds to an individual ping representing the location of a given truck at a single point in time. Table 1-1 demonstrates the fields associated with each ping in the dataset including anonymized carrier and truck IDs, latitude, longitude, date and time. In a given year, approximately 1.1 billion rows exist in the data file.
Table 1-1: Sample information accessible from raw GPS data

<table>
<thead>
<tr>
<th>Carrier ID</th>
<th>Truck ID</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Date</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1042</td>
<td>554</td>
<td>48.47848</td>
<td>-114.14864</td>
<td>20130302</td>
<td>145845</td>
</tr>
<tr>
<td>1042</td>
<td>589</td>
<td>52.54987</td>
<td>-108.13242</td>
<td>20130309</td>
<td>224532</td>
</tr>
<tr>
<td>1165</td>
<td>1147</td>
<td>47.34894</td>
<td>-109.78547</td>
<td>20130328</td>
<td>062234</td>
</tr>
</tbody>
</table>

1.7 Thesis Research Questions and Objectives

1.7.1 Mining truck GPS Big Data

A notable absence in transportation research utilizing GPS data is the creation of long distance inter-regional trips, yet GPS data is ideally suited towards this task for several reasons. First, the data does not require fixed detection resources since GPS transponders are already available on most trucks. Furthermore, the GPS data observes a given vehicle throughout the entire course of the journey. Therefore the data becomes economically advantageous for very large distances since permanent sensing infrastructure is not required.

Consider that at a local level, municipalities could purchase technology (such as Bluetooth) to install throughout areas of interest and track trips travelling across major areas of the city, but such an investment becomes infeasible for inter-regional vehicle movements where larger areas require an increasingly greater number of sensors. Moreover, the vehicles will have only been observed at the specific locations where sensors were installed, thereby losing the ability to identify the movement patterns at locations that are less busy (such as locations off the highway near the start/end of the trip). The latter makes identifying the origin and destination of the trips exceedingly difficult. To this end, the thesis seeks to address the following research question:
“Can opportunistic GPS Big Data be mined to produce reliable datasets capable of characterizing inter-regional and cross-border truck movements?”

This problem addresses a key gap in existing transportation data by creating a large dataset of origin and destination trips that cross municipal, provincial, and federal boundaries. Several challenges are tackled to meet this objective that are outlined in the sub-points below.

1.7.1.1 Devise processing methods to extract trips by origin and destination

Mining the raw GPS data for information is not a trivial endeavor since the collected records include approximately 1.1 billion records for each year, representing the movement of over 60,000 Canadian trucks across North America. While the dataset can be characterized by a large volume of information, each record contains only simple attributes on where a vehicle has travelled. The individual pings include identifiers for the specific truck and carrier as well as the latitude and longitude of the vehicle at a specific time. As such, the GPS pings can be sorted by time to track the movements of each truck.

Since the GPS data used in this thesis is stored in a database environment (Microsoft SQL), the necessary procedures are converted into scripts that can be utilized in the SQL environment to efficiently process the data into a database of trips. The methods of processing the GPS data are easily transferable to future research using this dataset or other sources of GPS data that will likely be utilized in the future. The scripts used to process the data (in Microsoft SQL) are included in the dissertation appendix for future reference. This information can be used to process future GPS datasets that are likely to be in high demand for transportation researchers in the near future as opportunistic data becomes increasingly available.
1.7.1.2 Identify useful patterns in the GPS data and devise geo-spatial algorithms to enrich the dataset

In addition to the general processing, the dissertation includes spatial data mining to infer information not explicitly by the opportunistic GPS data source. As a general guideline, the spatial component of the data is explored first in a GIS environment (Esri ArcGIS software). The devised procedures are then converted into the scripts from the previous objective to process the data efficiently in an SQL environment.

As one example, a method of identifying shipping depots for individual carriers is developed based on ping density. This provides an identification of the specific carrier (if needed) and can be used as an important start/end location for truck tour models (see Gingerich and Maoh, 2015). Moreover, validations of the locations based on Google Earth are performed to identify the name of the carrier and the location as a valid shipping depot.

A large focus of the processing in the thesis is placed on stop events. This includes the identification of stops for each truck as well as a classification of the type of stop. A definition is created to sort the stop events into two categories: primary stops where goods are transferred and secondary stops where other needs are fulfilled such as driver breaks and fuel refills. As part of the stop event classification, a new method is created by adapting Shannon entropy locations where stop events occur. The primary stops are specifically used in this dissertation as the end points for identifying trips from the GPS data since they represent events where goods are transferred.

The patterns identified in this dissertation provide valuable insights into mining truck GPS data. These patterns allow an analyst to derive useful information from the raw
data and create a richer processed dataset of trips. Moreover, the successful application of various patterns in the GPS data has broader implications towards the continued viability of opportunistic data for transportation research.

1.7.1.3 Devise an approach to correct potential bias in derived trip information

While the GPS data observes the movement patterns of a large number of trucks (approximately 60,000 per year), it represents a sample of all truck freight. For comparison, the full number of heavy trucks is estimated to be 317,219 (Natural Resources Canada, 2009). As a sample, the GPS data is prone to representation biases that will not accurately reflect the true pattern in commercial truck transportation. Two important biases are identified in the data. The first bias pertains to the distance travelled by the trucks. A comparison of the trip distances with those identified in the 2006 Ministry of Transportation Ontario (MTO) commercial vehicle survey (CVS) indicates that the GPS derived trips tend to traverse longer distances. Secondly, an analysis of the firms that are visited by trucks in the GPS dataset finds that certain industries are better represented (visited) compared to others. In particular, primary industries represented by mining and agriculture are not well represented.

A new method of expansion is proposed to create a dataset from the GPS data that matches total truck trips observed in Ontario while taking into consideration the aforementioned biases. This expansion includes trip generation, distribution, a shortest path network assignment, and a final optimization procedure.
1.7.2 Apply the GPS Dataset to Better Understand Cross-border and Inter-regional Truck Movements

The second major objective of the thesis is the application of the processed dataset towards a better understanding of inter-regional and cross-border truck movements. This objective serves to enhance the knowledge on long-distance Canadian truck movements as an under-represented area of transportation research. As such, the following research questions are posed:

“In a regional context:

1- What type of factors explain the border crossing choice of trucks moving between Canada and the U.S.?
2- To what extent does the border contribute to the total truck trip delay?”

GPS has certainly gained recognition as a viable dataset in recent years. As discussed earlier in the introduction, it has been previously utilized as a data source for truck tours, reliability measures, and trip routing problems. But such studies have only begun to explore the potential of GPS data as an immense source of trip information that can be used to solve numerous problems in academic research and more practical industry problems. For example, the data in this thesis has been utilized by the author to provide an Ontario company with optimized locations for natural gas fueling facilities to maximize coverage to trucking fleets.

The applications of the GPS dataset in this thesis presents a showcase of some of the potential that can be obtained at a macroscopic level (by aggregating trips) and a more detailed microscopic level (by assessing individual trips). Two unique applications of the GPS data are presented in this thesis as described in the next sub-points.
1.7.2.1 **Border crossing patterns and choices of trucks**

An extensive aspect of the research work is the application of GPS data towards understanding truck activity patterns between Canada and the U.S., with a particular focus on two of the highest volume border crossings in Canada: the Ambassador Bridge and Blue Water Bridge. These patterns are created based on the processed GPS data that creates information by combining the border information (crossing location, date/time, an wait time) with the origin/destination of each trip and an estimate of the industry associated with it.

Border crossing choice is also modelled in the thesis using a case study of trucks travelling between Toronto and Chicago. The processed GPS dataset highlights the volume of vehicles travelling between these locations. Moreover, The GPS trips are used in several different ways to estimate the border crossing choice model. First, the choice of individual trips derived from the GPS dataset are utilized as a dependent variable in the model. Secondly, aggregate values of temporal crossing time statistics derived from the GPS data are utilized (among other variables) as explanatory factors. As a result, the mined GPS data in this thesis is utilized in multiple forms to establish the border crossing choice model and highlights the usefulness of the processed GPS data at both an aggregate and individual level.

1.7.2.2 **Delays experienced by trucks between regions and countries**

The application of GPS data towards a better understanding of delays is also explored in the thesis. From an aggregate perspective, delay at the border is analyzed based on distributions of crossing time. While this is extremely useful information on its own, the thesis further explores the potential of GPS data by measuring how much of an
impact the delay at the border actually impacted the entire trip. This type of viewpoint takes advantage of the GPS derived trips that observe the entire journey of a trip instead of one single location. In addition, the dissertation provides a measure for delay that identifies the proportion of a trip affected by expected and unexpected delays. The distinction between the two is important since the costs of unexpected delays will be much higher compared to expected delays observed during the transfer of goods the supply chain process.

1.8 Thesis Outline

This dissertation concentrates on pushing the forefront of big data analysis in freight transportation by mining and processing opportunistic GPS truck data to create reliable transportation datasets, and utilizing the derived datasets to address pressing and timely truck transportation problems. To achieve the objectives set out in the previous section, the thesis follows a progression arranged to provide information on the raw data and required processing needs, create a new dataset of origin-destination trips with desired information, expand the dataset derived from sample data, and apply the dataset towards an analysis of cross-border truck movement patterns and characteristics of inter-regional trips.

Chapter 2 of the dissertation describes details of the GPS data, its characteristics, and the steps taken to process the information (objective 1.7.1.1). In addition, the methods and patterns utilized to add extra information to the data are discussed including the identification of shipping depots corresponding to individual carriers, the calculation of stop events, and classification of the stop events as primary or secondary (objective
1.7.1.2). This includes a new adaptation of Shannon entropy to assist in the detection of secondary stop events.

Chapter 3 discusses the final processing of the raw GPS data into trips (objective 1.7.1.1). The characteristics of the processed truck trips are then examined. This includes an analysis of international truck trips travelling between Canada and the U.S. across the Ambassador Bridge and Blue Water Bridge (objective 1.7.2.1).

Chapter 4 focuses on an important aspect of the GPS derived trips as a sample of all truck movements. This includes an identification of representative biases present in the GPS trips. An expansion approach is introduced to account for the bias deficiencies while also expanding the data to match totals observed in Ontario (objective 1.7.1.3).

Chapter 5 presents a logit model of border crossing choice for truck trips travelling internationally between Toronto, Ontario and Chicago, Illinois (objective 1.7.2.1). While several variations of routes exist, all trips travel across the Ambassador Bridge or Blue Water Bridge. Moreover, the average travel time for the trips observed in the dataset are relatively similar, providing an opportunity to determine other factors influencing the choice of border crossing.

Chapter 6 examines delays at the border to derive a distribution of crossing times and further relate the border delays to their overall effect on the entire duration of a trip. Moreover, delays observed in the trips are further deconstructed into a measure identifying the proportions of a trip experiencing expected delay and unexpected delay (objective 1.7.2.2).

Finally, Chapter 7 concludes the dissertation by describing the benefits of utilizing GPS data as demonstrated by the thesis. In addition, the conclusions highlight
the major results from the dissertation, their impact on the field of transportation, and finally discusses the current limitations and future directions of research. As a result of the thesis, new information is created that did not previously exist in transportation planning for Canada (and Canada-U.S. trade) and provides research that demonstrates the usefulness of data that already exists but is not currently utilized to its full potential. As data continues to be created at increasing rates by vehicles that are incorporating a greater reliance on technology (i.e. connected vehicles), this type of research will become increasingly valuable to the field of transportation and sets the stage for future work utilizing opportunistic data.

1.9 Chapter 1 References


*Computers, Environment and Urban Systems*, vol. 28, pp. 9–44.

Cambridge: *Cambridge University Press*.


quantity/delivery/vehicle-based model, *Transportation Research Part E: Logistics and 
Transportation Review*, vol. 65, pp. 84–98.

Assessing Temporal Effect of Economic Activity on Freight Volumes with Two-Period 

Publishing.


1126-1139.

origin destination using commodity based model combined with an empty trip model, 

Roberts, B., Rose, A., Heatwole, N., Wei, D., Avetisyan, M., Chan, O., Maya, I. (2014) 
The impact on the US economy of changes in wait times at ports of entry, *Transport 
Policy*, vol. 35, pp. 162-175.

Institute.

Sharman, B. (2014) Behavioural modelling of urban freight transportation: activity and 
inter-arrival duration models estimated using GPS data [PhD Thesis], University of 
Toronto.

Toledo, T., Sun, Y., Rosa, K., Ben-Akiva, M., Flanagan, K., Sanchez, R., Spissu, E. 
(2013) Decision making process and factors affecting truck routing, published in 
Freight Transport Modeling, eds: Ben-Akiva, M., Meersman, H., and van de Voorde, 
E., pp. 239-255.

Trading Economics (2017) Canada GDP [online website] Available at: 
<http://www.tradingeconomics.com/canada/gdp>

Transport Canada (2011) Transportation in Canada 2011 – Statistical Addendum, 
Available at: <https://www.tc.gc.ca/media/documents/policy/Stats-Addend-2011- 
eng.pdf>.


CHAPTER 2
DATA PROCESSING AND MINING

2.1 Introduction

Modern supply chain processes rely heavily on goods transferred between different manufacturers, distributors and retailers across various regions. While several modes of transportation can be used, the reliance on heavy trucks (discussed in Chapter 1) is very prominent in North America, particularly during the final segment of a supply chain. Increased freight activity, and higher complexity freight models, have prompted the need for detailed data to understand current processes and predict future freight travel demand. Previous data was often obtained through sampled surveys that can be expensive for the analyst and time consuming for the respondent (see Allen et al., 2012, for a comprehensive summary of survey types). An emerging alternative to surveyed data is passive information collected through Global Positioning Systems (GPS) technology.

Since the late 1990s, GPS transponders have been frequently used by freight carriers to track the current position of truck fleets. However, the usage of GPS truck data in freight analysis has only become more commonplace in recent years. For example, Allen et al. (2012) observe that only 3 out of 284 studies pertaining to urban freight transportation examined between 1960 and 2008 used GPS data. The technology has been gaining increased attention for research purposes due to the increasing availability of GPS data for research purposes. For example, several studies in the U.S. have been published that utilize
truck GPS data derived from the American Transportation Research Institute - ATRI (see for example: Liao, 2014; Kuppam et al. 2014; Bernardin et al., 2015).

GPS data provides enormous potential for understanding the current patterns of freight based on the large volume of data alone. However, the volume of data also presents a substantial challenge since it was not originally intended as an input for transportation models and analysis. In essence, the raw GPS data tells us where trucks have travelled but it does not provide any detailed information on what is happening or why it is occurring. For example, the purpose of truck movements and stops observed with GPS data is initially unknown unless accompanied by travel diaries that provide ancillary information. Therefore a need exists for novel methods to mine this data in order to understand various patterns given that ancillary information is not typically collected.

The rest of Chapter 2 will be presented as follows. A description of the raw GPS data used in this thesis is first provided along with general data statistics. Next, a method of identifying the location of shipping depots is conceptualized in ArcGIS software before implementation in Microsoft SQL where data can be processed more efficiently. The shipping depots serve to identify the type of carriers that make up the dataset and the spatial distribution of their home base of operations (i.e. where the trucks typically reside when not in use and the starting point for many truck tours). Following the analysis of shipping depots in the dataset, a substantial portion of this chapter is focused on stop events including (i) identifying stop events in the GPS dataset, (ii) classifying the purpose of each stop event, and (iii) associating an industry type to each stop event. The processed stop events are important features that will be used in Chapter 3 to identify truck trips. Finally, the overall approach used to process the raw GPS data is presented at the end of the chapter.
2.2 GPS Data Characteristics

As discussed in Chapter 1, the GPS data obtained for this thesis is a form of opportunistic data that was created for a different purpose. More specifically, this GPS data was originally created by trucks belonging to carriers that subscribe to the tracking services provided by Shaw Tracking. As shown in Figure 2-1, the GPS data is originally created as a method of tracking the movement of the trucks in real time.

![Figure 2-1: Original purpose of GPS data](Source: Spec India, 2017)

2.2.1 GPS Data Attributes

Once the GPS data is created and used for its original purpose, it typically sits unused in storage. In this case, the data is transferred from Shaw Tracking to researchers interested in utilizing the data for transportation purposes. As shown previously in Table 1-1, the data in its original raw form contains records (rows) that correspond to individual GPS pings. Each ping provides truck movement information by location (i.e. latitude and longitude) and time. Numerical identifiers per ping are also provided to differentiate consistently track individual trucks and carriers. Sorting the data by time for a given truck provides an opportunity to observe the travel patterns of the vehicles.
Trucks in the dataset travel across both Canada and the U.S. However, this particular GPS data source corresponds specifically to Canadian owned trucks. The data is not provided in real-time, therefore the processing and analyses in this thesis are based on past events. The timeline of the GPS data includes September 2012 to December 2014 with approximately 1.1 billion pings per year. This dataset was obtained as part of a collaboration with Transport Canada and is no longer available for use (in its original raw form) following the completion of the project. However, more recent data in 2016 was purchased directly from Shaw Tracking with one month currently processed for March, 2016.

The interval between GPS pings can vary substantially, from single seconds to an hour or more, but an interval of 5 to 15 minutes is typically observed when the vehicle is moving. This is notably different from some other GPS data sources with more frequent pings. The reason for the larger intervals is not explicitly known, but the GPS data provider may utilize sparser pings to reduce their data input velocity or delete some pings in storage to reduce the data size.

2.2.2 GPS Data Accuracy

In terms of accuracy, it can be safely assumed that the ID fields for truck and carrier are consistently correct. However, the truck ID field is only unique for a given carrier and may be repeated by other carriers. Therefore unique ID values are created by concatenating the carrier and truck ID fields together.

The time stamp for each GPS ping is provided to the nearest second. While the time stamp is very likely to be correct, a minor error will not affect the data as long as the time
shift is consistent across all pings (i.e. the observed patterns will look the same even if all
points are shifted in time by a small amount).

The largest culprit of potential accuracy issues arise from the location attributes. The latitude and longitude are typically received in this dataset with a precision of 5
decimal places (such as a latitude of 75.39586). The coordinate system used by the GPS
devices is the 1984 World Geodetic System (WGS), where the last decimal place represents
approximately 1.1 meters, although the length of longitude varies based on location. While
some GPS devices can achieve accuracies on par or better than this precision, it typically
comes from relatively advanced methods such as differential GPS computations requiring
multiple receivers at different locations (see Trimble, 2017 for more information). In
reality, the dataset used in this thesis achieves a lower accuracy than the precision allows.
For example, Figure 2-2 below shows GPS pings at the Ambassador Bridge for a one
month period. The dense line (formed by a large number of pings) represents the location
of the bridge while the sparser points to the left and right represent noticeable location
errors since they occur directly above water.

![Figure 2-2: GPS ping dispersion at the Ambassador Bridge](image)

35
Although one could rely on the dilution of precision (DOP) measurement to assess the quality of the captured GPS readings (see Langley, 1999, for more information on DOP), this information is not included with this dataset. As an alternative, an analysis of lateral errors (with respect to the direction of travel) was conducted on two sections of road including the Ambassador Bridge (in Figure 2-2) and a section of Highway 401 on the Canadian side (between exits 28 and 34 as shown in Figure 2-3). Errors along the direction of travel cannot be measured since they are indistinguishable from correct locations. However, this analysis measures errors along both axes since the Ambassador Bridge is oriented roughly north/south and the section of Highway 401 is oriented east/west.

Figure 2-3: GPS ping dispersion along a Highway 401 corridor

The measured maximum errors are 439 meters and 593 meters at the Ambassador Bridge and Highway 401 section, respectively. However, large errors are fairly rare in comparison to the total number of observed pings at each location as shown by the average errors of 15.4 meters and 27.7 meters, respectively (Table 2-1).
Table 2-1: Lateral GPS errors

<table>
<thead>
<tr>
<th>Location</th>
<th>Count</th>
<th>Avg.</th>
<th>St. Dev.</th>
<th>95 %</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ambassador Bridge</td>
<td>11,843</td>
<td>15.4 m</td>
<td>44.4 m</td>
<td>89.6 m</td>
<td>439.3 m</td>
</tr>
<tr>
<td>Section of Hwy 401</td>
<td>10,602</td>
<td>27.7 m</td>
<td>58.8 m</td>
<td>144.9 m</td>
<td>593.2 m</td>
</tr>
</tbody>
</table>

2.3 Identifying Carrier Shipping Depots

One of the major types of locations visited by trucks is the shipping depot. These truck yards serve as a home base where trucks for a given carrier may be stored between trips. From a practical perspective, these areas are important because they comprise a large proportion of trip activity productions in our sample compared to areas with no shipping depots. Moreover, these depots provide a location to delineate truck tours and/or trips that begin and end at the shipping yard. Finally, the type of carrier can be determined after identifying their corresponding depot. While the individual carriers are not the focus of the thesis, their identities provide valuable information on the type of companies in the dataset.

Identifying the location of the main shipping depot for each carrier was performed on the basis of the following two assumptions:

1. The depot will occur at a location that is frequently visited by trucks belonging to the given carrier

2. The first GPS ping obtained from each truck is more likely to occur at the shipping depot compared to later GPS pings

This section of the thesis used data pertaining to March, 2013. However, the beginning date used to select the GPS pings may have an impact on the estimated location of shipping depots. For example, using a dataset that starts in January (or directly after a
holiday) may result in fewer occurrences of truncation where the first observed point is actually occurring in the middle of a trip.

### 2.3.1 Identifying Shipping Depots Using Clustering

The result of the above assumptions is that a method of clustering is needed to take the first GPS ping from each truck and determine the location with the highest occurrence. Various clustering methods already exist to perform this task. For example, a study by Sharman and Roorda (2011) evaluated different partitioning and hierarchical agglomerations clustering techniques to determine trip end locations from GPS data. Another viable technique is kernel density estimation, which can be used to evaluate the locations where point events are highly clustered over space (Bailey and Gatrell, 1995).

The above kernel density method is implemented by creating a density surface across space consisting of the first point from each truck belonging to a given carrier. An illustration of the points and resulting kernel density cloud in (calculated in ArcGIS\(^1\)) are provided in Figure 2-4. However, the processing time using such software can increase dramatically due to the size of the dataset, size of the study area stretching across Canada and the U.S., and the required precision for a final location. For the latter, a higher precision leads to smaller raster cells, which subsequently increases the computation time. As a result, a computationally quick method of clustering in Microsoft SQL was devised and implemented. A description of this approach is provided next in Section 2.3.2 while a comparison of the computation times and results for the two approaches is provided afterwards in Section 2.3.3.

---

\(^1\) ArcGIS is a mainstream Geographic Information System software used to process and represent spatial information (ESRI, 2013)
2.3.2 Identifying Shipping Depots Using Microsoft SQL Server

An alternative approach was devised to determine the shipping depot using Microsoft SQL Server 2008; thereby reducing the processing time and eliminating the need to transfer data to GIS software at this stage of processing. For a given first ping \( t \) of truck \( t \) belonging to carrier \( c \), whose latitude and longitude coordinates (in decimal degrees) are \( Y_{ctf} \) and \( X_{ctf} \), respectively, the clustering procedure begins by multiplying the coordinates to provide a single value \( (Z_{ctf}) \).

\[
Z_{ctf} = X_{ctf}Y_{ctf} \tag{Eq. 2-1}
\]

Next, the coordinate \( Z_{ctf} \) is rounded to \( p \) significant digits. The rounding procedure is used to group nearby points together. A lower value of precision \( (p) \) results in a larger area that is used to bind the points into a cluster. The resulting coordinate \( (Z'_{ctf}) \) is used as an identifier for the points located in a given cluster.

\[
Z'_{ctf} = \text{ROUND}(Z_{ctf}, p) \tag{Eq. 2-2}
\]

The \( Z'_{ctf} \) calculated here is not necessarily unique and may occur for several different locations occurring across an arc over space. For example, consider that a shift in a given point with an increase in latitude \( (Y_{ctf}) \) and a similar decrease in longitude \( (X_{ctf}) \) will result in the same \( Z'_{ctf} \). As such, an arc carrying the same \( Z'_{ctf} \) exhibits a northwest-
southeast directionality. The risk of obtaining the $Z$ value at multiple locations is mitigated by rounding to a larger number of significant digits ($p$), since this reduces the width of the arc (and corresponding size of the location clusters). But this does not remove the possibility of non-unique $Z$ values entirely. Therefore a constraint is added below to ensure that all points for a given cluster have a similar latitude and longitude. However, future applications of this type of clustering should be carried out with caution due to the non-unique nature of the resulting values.

For a given carrier $c$, a numeric variable $Z'_{c}$ is calculated as the mode of $Z'_{ctf}$ from all trucks belonging to the carrier to determine the pings that pertain to the densest location.

$$Z'_{c} = \text{MODE}(Z'_{ctf} \in c) \quad \text{(Eq. 2-3)}$$

The latitude and longitude coordinates of the shipping depot for each carrier $c$ are obtained by averaging the location of each point belonging to $Z'_{c}$.

$$X_{c} = \text{AVERAGE}(X_{ctf} \in Z'_{c}) \quad \text{(Eq. 2-4)}$$

$$Y_{c} = \text{AVERAGE}(Y_{ctf} \in Z'_{c}) \quad \text{(Eq. 2-5)}$$

Due to the non-unique nature of the resulting $Z$ values, a spatial constraint is implemented in Equations 2-4 and 2-5 (and retroactively to Equation 2-3) to ensure that points pertaining to only one location are included in the calculations. This is performed by filtering the X and Y coordinates independently such that point events are discarded if they are not located near the densest location. As a result, if multiple cluster locations exist with the same values of $Z$, the cluster with the largest number of points is retained since the algorithm is only interested in determining the densest location (corresponding to the two assumptions listed earlier at the beginning of Section 2.3).
2.3.3 Shipping Depot Identification Results

The algorithm that was executed in an SQL environment identified 475 shipping depots out of 750 carriers for a one-month period in March 2013. The remaining carriers did not have a cluster with 3 or more points to properly identify a shipping depot. This becomes more common when a given carrier has fewer trucks observed in the dataset, however, setting the date to begin at the start of January (and extending it for a full year) may help improve the number of results. While the calculations in ArcGIS (Section 2.3.1) had a runtime longer than 6 days, the SQL algorithm (Section 2.3.2) produced results in 5 seconds on the same computer (Pentium i7-core 3.4 Ghz, 32 GB Ram). An aggregate count of shipping depots by Canadian census division is given in Figure 2-5, however, the exact location of each depot is not provided to protect the privacy of the corresponding firms. The Region of Peel in the Greater Toronto Area has the highest number of primary truck yards in this dataset with 39 observations.
2.3.4 Validations of Results

The results from the SQL-based clustering procedure were validated using the kernel density estimation results from ArcGIS as the true representative of the shipping yard locations. The kernel density results provided more locations than the SQL algorithm (much closer to the total of 750 carriers), but this was caused by a lack of constraint with regards to the number of points (or density size) needed to provide a reasonable indication of the truck yard.

There were 50 outliers identified out of the 475 locations where both methods were able to identify a location. These outliers represent carriers where the two methods did not locate the shipping depot in the same geographic vicinity. A manual verification of these 50 points revealed that these errors were due to the occurrence of overlapping clusters that indicate the presence of multiple shipping depots. Accordingly, future uses of this technique could be adjusted to consider the possibility of multiple truck yards. A statistical analysis of the two methods, which excluded the 50 outliers, revealed a root mean square error (RMSE) of 70 meters.

A second validation was performed to test whether the results actually correspond to the spatial location of a carrier shipping yard. A 10% random sample (50 points) of the identified shipping depots was drawn and assessed using Google Maps and Google Street View. The analysis found that all of the locations resulting from the SQL clustering approach proposed here are correctly located at a carrier’s shipping depot. Moreover, periodic spot checks of various shipping depot locations (other than the selected 10% random sample) have shown that the locations are shipping depots in most cases. Locations
that were not identified as shipping depots were gas stations that are frequently utilized by a given carrier.

The results from the validations suggest that the original two assumptions generally hold true: (1) the shipping depot is frequently visited by the trucks corresponding to the carrier and (2) the first ping from each truck for a carrier can be used to identify these locations. Other data could be used to contribute towards the identification of shipping depots. This is particularly relevant if the analysis expands to identify multiple shipping depots for a given carrier. An example of a potentially useful variable is the dwell time of stop events at the locations identified as possible shipping depots (stop events are discussed next in Section 2.4). For example, some trucks may be stored at the shipping depot for relatively long periods of time in comparison to other stop events where the vehicle is still completing the intended movement of goods. Moreover, entropy (as will be discussed later in Sections 2.5.3 to 2.5.5) related to locations with a large variety of carriers has been shown in Figure 2-10 to be lower for shipping depots compared to other stop locations since the corresponding carrier tends to be the primary user of the facility.

2.4 Stop Events

The time stamp information provided for each GPS data record is used to calculate the elapsed time between successive GPS pings and subsequent dwell time that accumulates if the truck is stopped. The calculation of the elapsed trip time is relatively trivial as the difference in time between two consecutive pings. However, the identification of a purposely stopped truck requires more attention. To determine if a vehicle is stopped, a number of previous studies have utilized travel speed as derived from the observed distance and time of consecutive pings (see for example: Du and Aultman-Hall, 2007;
Gong et al., 2012; Kuppam et al., 2014; Yang et al., 2014). However, such a metric may result in erroneous stop identification for vehicles moving at low speeds due to congested conditions. In addition, the speed can be susceptible to ‘signal jiggle’ where the GPS location bounces to an incorrect position and results in a velocity that is artificially high.

As an alternative method for identifying stop events, a distance measurement was devised as shown in Figure 2-6. After the pings for a given truck are sorted sequentially according to the registered time stamp, the location of a first ping (P₁) is compared to the location of the next ping (P₂). If the distance is less than a certain threshold \( l \), the dwell time \( (d_w) \) is set equal to the elapsed time between the two pings. If the next ping (P₃) is also less than a distance \( l \) from the first ping, the dwell time continues to accumulate. This continues until there is a ping (say Pₙ) located outside the buffer threshold at which point the dwell time is reset.

**Figure 2-6: Distance based dwell time calculation**

The buffer threshold used to determine a stopped vehicle is set to a radius of \( l = 250 \) meters in this thesis. The radius is imposed to avoid cutting short a stop event if a vehicle moved a limited range within a given property. In such a case, it is often more convenient to assume that the ending of a stop event occurs when the vehicle leaves the property location. The chosen buffer also accommodates any spatial errors that might arise due to bad GPS readings. As seen in Table 2-1, the 95 percentile error for GPS pings was
found to be 89.6 m and 144.9 m. In addition, it was found that only 1.6% and 1.8% of the pings at the two sites exceeded an error of 250 meters. Therefore, the majority of errors are unlikely to exceed the chosen buffer radius.

Using the above approach, there are approximately 2.7 million observed stops in the analyzed GPS dataset for one month in March, 2013. Each stop was identified when a truck had a total dwell time exceeding 15 minutes. This value may seem large for a person familiar with short urban deliveries. However, this limit is considered suitable for the longer distance trips that characterize truck movements in this dataset and the focus of the thesis on inter-regional and cross-border trips. In addition, a larger minimum dwell time helps reduce the possibility of false positive stop events (in the sense that the vehicle did not stop for any intended purpose) that may occur under severe congested traffic conditions. This would pertain to vehicles travelling with an average speed below 1 km/hr based on a 250 meter threshold and 15 minute minimum dwell time. Unless complete gridlock occurs, this is not a likely scenario even in congested situations.

To investigate the potential of false positives from congested traffic, an examination was conducted for an area containing Highway 401 road links between Highway 407 and Highway 403/410 in Ontario, Canada (latitude between 43.588 and 43.638; longitude between -79.819 and -79.661). This area was selected since it has the largest concentration of GPS pings in the dataset and occurs along a heavily congested highway corridor in the Toronto metropolitan area in Ontario. Only 48 out of 32,174 stop events in the examined area are false positive stop events based on observed events occurring directly on the road, suggesting that the potential for obtaining erroneous results are kept to a minimum with the proposed techniques.
2.5 Classifying Stop Events

2.5.1 Definitions for Stop Event Classification

Although stops are observed based on the approach provided in the previous section, their purpose is not explicitly provided due to the lack of ancillary information on record. One of the most important events for freight movements occurs when the vehicle is stationary. Stop events are classified in this thesis as one of the following:

1. Primary stop event, which occurs when goods are transferred between the truck and location (or another truck)

2. Secondary stop event, which occurs when a truck is stationary for other purposes such as driver breaks or fuel refills

Primary stops are particularly important for transportation models since they denote what would typically be considered a trip end for a given truck. Likewise, secondary stops are useful as a complement to primary stops by providing a complete picture on the nature of truck movements over space. However, it is important to make the distinction between the two types of stops since they correspond to different activities. Any transportation model that utilizes micro-data (such as the GPS based information) on truck movements for its calibration will require an accurate representation of primary and secondary stops to properly represent travel activity patterns. The trip ends in this thesis correspond to primary stops since they determine the origin and destination of goods that are shipped. Errors occur if secondary stops are misclassified as possible trip ends since goods are not actually transferred. Consequently, failing to classify truck stop purpose will lead to a miscalculation of the actual trip end locations.
2.5.2 Previous Approaches to Stop Event Classification

The dwell time of a truck stopped at a given location can provide information on the nature of the stop event in some cases. Hess et al. (2015) derived intervals of dwell time to differentiate stop types. Dwell times less than 2 minutes are disregarded as too short to conduct a valid stop. A dwell time interval of 2 minutes to 15 minutes is assumed to be a primary stop on the basis that secondary stops require a duration exceeding 15 minutes. A dwell time greater than 45 minutes is assumed to represent the end of a daily delivery schedule since this exceeds the maximum break time for daily trips in the European Union (EU).

Such rules are based on strong assumptions derived from regulations for short trip drivers in the EU which would cause issues with the dataset used here. The latter GPS dataset used in this thesis is derived from carriers that perform a large proportion of long-haul trips that potentially exceed one day (this was confirmed by the carriers identified by shipping depots in Section 2.3). Therefore the upper 45 minute threshold is not applicable. Moreover, a 2 minute delivery time may be suitable for very short urban stops but would be too low for heavy trucks with larger deliveries. The short dwell time also raises a potential issue of false stops occurring due to congestion on roadways experiencing heavy traffic. Regardless, Hess et al. (2015) still require additional methods for differentiating primary and secondary stops when there is a dwell time interval between 15 to 45 minutes. In such a case, a database of locations is used to identify secondary stop events.

Bohte and Maat (2009) also used proximity to known locations of rest stops and gas stations to identify secondary stops in their study of passenger GPS data. However, a comprehensive list of all secondary stop locations becomes increasingly difficult to obtain.
as the size of the study area and the number of political boundaries is increased. Land use and land cover data sources are generally available but not detailed enough to identify secondary stop locations. Parcel level data with information on firm/industry type could provide enough information for the latter but is not readily accessible. Municipal governments typically retain such information but it can be difficult to access due to the sheer number of municipal governments (this thesis covers all of Canada and the U.S.) and their reluctance to provide detailed parcel information based on confidentiality concerns. Parcel data boundaries can also be useful for clustering stop events as evidenced by Sharman and Roorda (2011).

Another option to determine the purpose of stop events is the utilization of datasets containing individual firms by location and industry. Free data is available using sources such as Google, Factual, and YellowPages, who include Application Program Interfaces (e.g. Google API) to connect to their databases using custom programs. However, an analysis found that they only represent a small proportion of total firms, leading to poor results for location identification. Other sources of firm data (such as InfoCanada) are available but require expensive annual fees and still have occasional data gaps and inaccuracies. The utilization of firm databases as an extra source of information in this thesis is described in more detail in Section 2.6.

Other methods of identifying the purpose of stop events have also been utilized in the past. For example, Yang et al. (2014) trained a model using machine learning algorithms to identify correct stops of urban grocery store deliveries in New York City. Their algorithm primarily utilizes three variables: Dwell time (stop duration), distance to the center of Manhattan, and distance to the nearest bottleneck area (congestion points).
However, this model was based on only 42 deliveries at limited locations. The use of machine learning algorithms has certain drawbacks including a low adaptability to circumstances that are different from the trained data (such as various industries and locales). Moreover, an extensive set of information that contains the actual purpose of a stop is required to properly train the model.

A model comprising multiple variables was calibrated by Du and Aultman-Hall (2007) to identify valid stops constituting trip ends for passenger vehicles. Their work employs a combination of three variables to correctly identify trip ends including the dwell time, distance to the road network, and heading change. While this is not directly related to the identification of primary and secondary stops for trucks, the heading change variable could be adapted in the case of truck stops. When dealing with trucks, a small heading change suggests the vehicle performed a secondary stop since this change occurs as a convenience along the route to the primary stop. However, the heading change variable could be influenced by the topology of the road (e.g. a bend in the road) leading to false classifications.

All of the potential methods listed above are based on an analysis of each stop event in isolation from each other. However, this thesis proposes that the large volume of information contained in the opportunistic GPS dataset can be leveraged by evaluating the pattern that emerges from analyzing stop events over space. As discussed in the next section, the entropy technique adapted here is well suited towards GPS data since the location of each stop event and the associated truck carrier identifier is the only information required.
2.5.3 Entropy Method for Classifying Stop Events

When analyzing the GPS data, it is expected that secondary stop locations attract a greater variety of trucks belonging to different carriers compared to primary stop destinations since the former provide useful services to any truck passing nearby. This diversity of carriers dwelling at a stop location can be captured by associating a larger variety of carriers with a higher level of entropy. As such, entropy is an ideal concept for classifying the type of stops. To the author’s knowledge, this is the first study to mine truck GPS data through the entropy technique in an attempt to identify the purpose of stopped truck events.

The application of entropy to classify stops is tested on a large dataset composed of 100 million GPS pings, which occurred during the month of March, 2013. The data, which is collected by Shaw Tracking, contains 40,650 individual trucks that belong to 750 Canadian owned carriers for this specific month. While the Canadian ownership results in a larger proportion of observed movements across Canada, many trips cross the Canadian border to the U.S. as a result of the international trade with Canada’s largest trading partner. For example, there were approximately 17,500 occurrences of trips in one month (March 2013) identified moving through the busiest Canada-U.S. border crossing at the Ambassador Bridge (see Chapter 3 for more details).

Entropy is a well-known principle that has been used in various disciplines to describe the state of order or chaos of a given system. Although it was first used in the natural sciences, entropy was also introduced in the field of information science in the late 1940s to analyze the amount of information embedded in transmitted messages (Shannon, 1948). Since then entropy has been successfully applied in other areas including
transportation and land use. For instance, Wilson (1970) applied the physical concept of entropy to model the flow of trips (i.e. spatial interaction) within a transportation system. On the other hand, Cervero (1989) utilized the Shannon entropy formulation to evaluate the level of land use mix in suburban neighborhoods. For a given neighborhood \( i \), an entropy measure \( EI_i \) is calculated using the following formula:

\[
EI_i = -\frac{\sum_{k=1}^{K} p_k \ln(p_k)}{\ln(K)}
\]  

(Eq. 2-6)

where \( p_k \) is the proportion of the area pertaining to land use type \( k \) from the total area of neighborhood \( i \), and \( K \) is the total number of land use categories in neighborhood \( i \) (Maoh and Tang, 2012). In the above formula, \( EI_i \) will take on values between 0 and 1 since the formula is scaled by the denominator \( \ln(K) \). \( EI_i \) values closer to 0 suggest land use homogeneity in neighborhood \( i \), while larger \( EI_i \) values approaching 1 denote a greater mix of land use types.

An entropy index \( EI_q \) similar to the one listed in Equation 2-6 is created here to quantify the variety of carrier fleets using a particular stop location \( q \) as follows:

\[
EI_q = -\sum_{c=1}^{C} \left( \frac{n_c}{N} \right) \ln \left( \frac{n_c}{N} \right)
\]  

(Eq. 2-7)

where \( n_c \) is the number of truck stop events occurring for a given carrier \( c \) at location \( q \), \( N \) is the total number truck stop events at \( q \) and \( C \) is the total number of carriers associated with all stopped trucks at location \( q \). More carriers with trucks stopping at location \( q \) will result in a larger entropy value. Conversely, a location with stopped trucks belonging to only one particular carrier will have an entropy value of 0 as shown in Figure 2-7. It is expected that primary stop locations, where the loading/unloading of goods takes place, exhibit fewer carriers and lower entropy. This is also true for carrier shipping depots where
the trucks for a given carrier or several carriers reside when not in use. Therefore, the locations with lower entropy (i.e. lower variety of carriers) will be less likely to provide a secondary function compared to stop locations with higher entropy.

**Figure 2-7: Entropy index (EI) interpretations**

Unlike Equation 2-6, the revised entropy formula in Equation 2-7 does not have a denominator \( \ln(C) \). As such, \( EI_q \) is not bounded by an upper limit of 1 as in the case of \( EI_t \). This treatment is critical for the classification of stopped truck events in this thesis. Our ability to classify secondary stop locations is dependent on identifying those locations that are visited by a large number of trucks from a large variety of carriers. Restricting \( EI_q \) by an upper bound of 1 will not allow us to correctly identify secondary stop locations. Consider one location visited by only two trucks belonging to two carriers and another visited by 1000 trucks belonging to 1000 carriers. With the denominator in Equation 2-6, both sites will have the same entropy value of 1. However, the 1000 carrier location is more likely to represent a secondary facility compared to the two carrier location, all other things being equal. This is better represented by the \( EI_q \) entropy in Equation 2-7 where no denominator is included.
The potential impact from the volume of carriers on the entropy values obtained by Equation 2-7 is demonstrated in Figure 2-8, where the highest potential $EI_q$ value increases as more carriers utilize the location. Moreover, the maximum entropy value occurs when the proportion of stops, $\frac{n_c}{N}$, is equal among each carrier $c$. For instance, consider two carriers (1 and 2) with the total number of stopped truck events $N$ constrained to 100 (note here $N = n_1 + n_2$). The entropy $EI_q$ can be calculated for different $n_1$ and $n_2$ combinations, as shown in Figure 2-9. According to the given parabolic curve, the maximum entropy value of 0.693 occurs when the proportion of stopped truck events is split evenly among the two carriers (i.e. $n_1 = n_2 = 50$). In summary, both the volume and variety of carriers with trucks stopping at a given location will increase the entropy value. These conditions are expected to occur more frequently for secondary stop locations where trucks will visit for various non-primary needs and provisions.

![Figure 2-8: Maximum entropy values by number of truck carriers](image-url)
2.5.4 Entropy Validation Results

To apply the concept of entropy, the locations of stops where trucks dwelled for over 15 minutes are sought. Stop events are clustered together to identify the locations where higher levels of stop events are observed. Locations with very few stop events are not desired for the analysis of entropy since they will automatically be classified as primary stops based on the formulation of the entropy index as shown in Equation 2-7. A brief discussion on clustering was provided previously in Section 2.3.1. Moreover, the method utilized here adopts the approach described in Section 2.3.2 where clustering is performed in Microsoft SQL Server. The result is a total of 3,370 clusters (i.e. stop locations $q$) which were identified throughout Canada and the U.S.
It should be noted that an alternative to clustering is the calculation of entropy for each individual stop event. In such a case, entropy is calculated for each stop by capturing nearby stop events with a predefined radius. This method was not used here since the distribution attained from the results would be weighted by the number of stop events at each location (instead of one observation per location). However, this approach for individual stop events is applied in the processing workflow of the GPS described in Section 2.7. Given the very large number of stop events, a test was performed on a random sample of individual stop events and the results were compared to the calculated cluster-based entropy values. A comparison of the results found very similar patterns between the two approaches.

The entropy given by Equation 2-7 was applied to each of the identified clusters in addition to shipping depots identified in Section 2.3. The latter are expected to have low values of entropy since each depot will be predominantly visited by the carrier to which it belongs. A histogram of the entropy variable $EI_q$ for the 3,370 clusters has values ranging from 0 to 4.78, as depicted in Figure 2-10. An interesting outlier in the histogram is the large presence of clusters with $EI_q$ values between 0.6 and 0.8. To understand this phenomenon, one can utilize the information given in the parabolic curve shown in Figure 2-9. The latter demonstrates that a maximum value of 0.693 exists when the proportion of stopped truck events is equal for a scenario involving two carriers. An exploration of the results finds that the two carrier scenario makes up the majority of the clusters producing the spike in Figure 2-10. More specifically, out of the 494 total clusters containing stopped truck events with an $EI_q$ value rounded to 0.7, there are 396 (i.e. 80 percent) clusters containing only two carriers.
Figure 2-10: Histogram of entropy results for all clusters and shipping depots

It is expected that shipping depots are typically utilized by a small number of trucking firms that own or lease the property. This assertion is supported by the results in Figure 2-10, which shows that the entropy for shipping depots are on average lower compared to all stop clusters with a mean of 0.85. The distribution of entropy between 0 and 1 for the shipping depots mirrors the distribution of all stop clusters. However, the proportion of shipping depots with an entropy exceeding 1 is considerably low. By contrast, this is not the case for the majority of stops.

Another *a priori* expectation for the entropy index was that high entropy would correspond to stops used for secondary purposes. The 150 locations with the highest entropy (*Elq* ranging from 3.97 to 4.78) were manually validated using Google Maps and Google Street View to determine the type of stop. A total of 148 out of the 150 clusters correspond to secondary stops such as truck stops, gas stations, and several motels. This initial validation establishes that high entropy at a given location is indicative of a secondary stop, thus supporting our postulated hypothesis.
A further exploration of the entropy characteristics was performed by considering a random sample of 250 clusters. This random sample is representative of the full population of stop clusters with a 94% correlation when comparing the frequency of clusters within entropy bins with intervals of 0.2. The exploration started by classifying the clusters as secondary or primary based on aerial views provided by Google Maps (Google, 2017). Figure 2-11 presents the results of this validation. Between an $EI_q$ interval of 0 to 1.6, the proportion of secondary stops fluctuates, but increases afterwards up to an $EI_q$ value of 2.8. The prevalence of secondary stops becomes more evident for locations with entropy index ($EI_q$) values exceeding 2.8.

Figure 2-11: Stacked bar chart of stop purposes by value of entropy

A transition period exists for clusters with $EI_q$ values between 2.5 and 3 in which the proportion of secondary stops to other stops begins to noticeably increase. This observation raises the question of what specific threshold value should be considered to separate secondary stops from primary stops. To address this issue, an additional 150 randomly selected clusters with $EI_q$ values corresponding to the transition period were drawn and classified as primary or secondary.
A threshold value of 2.75 was selected based on a noticeable separation in the proportion of secondary stops as shown in Figure 2-12. These results indicate that there is a large increase in the proportion of secondary stops from 80% to 88% for the \(EI_g\) ranges of 2.5 - 2.75 and 2.75 - 3, respectively. The threshold value of 2.75 provides a high confidence that the stop location is used for secondary purposes. As seen in Figure 2-13, 95.8% of all locations with entropy values above 2.75 were validated to be secondary stops. This leaves a reasonably small percentage of type I false positive errors at 4.2% that represent primary locations mislabeled as secondary locations. While this is not explored further in this thesis, it would be beneficial to further investigate these primary locations to determine why they exhibited an uncharacteristically high entropy for their stop purpose.

**Figure 2-12: Stop purpose for entropy ranging from 2.5 to 3 (150 point sample)**

**Figure 2-13: Stop purpose for full entropy range (250 point sample)**
While the right leaning threshold value of 2.75 keeps type I errors minimal (i.e. 4%), this leaves a large number of type II errors in which secondary stops are falsely identified as primary for entropy less than 2.75. As suggested by Figure 2-13, 29% of the stop locations with entropy below the 2.75 threshold are actually secondary stops. The type II errors reveal a sub-optimal solution if this method marked the completion of the data processing. However, further analysis of the clusters and their associated entropy was performed to identify an optimal threshold that could minimize the severity of type II errors in the dataset.

Understandably, there is some equilibrium between type I and type II errors that will depend on the chosen threshold values. The importance of each error type to the analyst will therefore have some impact on the selection of these thresholds. For instance, a cut-off $EI_q$ value of 3.2 will result in zero type I errors since all locations with entropy above this value were found to be secondary stops. However, type II errors will represent 56% of the locations with entropy below 3.2. Shifting the threshold value of $EI_q$ down to 1.8 will establish equivalence between type I and type II errors in which 19% of stops in either case will be incorrectly classified.

2.5.5 Entropy Method Discussion

The entropy estimation discussed in this thesis was tested using passively collected GPS pings for one month of truck movements in March, 2013. The analysis of 101.6 million individual GPS pings provides a substantial challenge but also affords opportunities to study spatial patterns that are not available when using smaller datasets. A novel approach is developed here to help evaluate the purpose of an observed stop occurring for a given truck. A larger value of entropy was postulated to occur for secondary
stops due to the utilitarian purpose of these locations. The analysis found that 148 of the 150 clusters with the largest entropy values are secondary locations. Moreover, a validation conducted on a random sample of stops found that 96% of the locations with entropy values above 2.75 provide secondary functions.

Due to the high percentage of secondary stops exhibiting an entropy value above the 2.75 threshold, it would be practical to automatically classify stops with high entropy as secondary (this is subsequently adopted as part of the GPS processing in this thesis as seen later in Section 2.7). However, doing so would not result in the identification of all secondary stops. Therefore it is suggested that this method should be used in coordination with other rules to identify secondary stops exhibit a lower entropy (see Section 2.5.5 for brief details of the other methods used).

Overall, the entropy validation results suggest the existence of a strong connection between the purpose of a stop and the variety of carriers that utilize the dwelled location. As such, the devised approach can be employed to process large GPS truck data for use in the development of freight transportation models. Naturally, the ability to derive the purpose of a stopped truck provides pivotal information for the understanding of vehicle movements that is typically not provided with passive GPS data.

Arguably, a limitation to the validation here is that it was only conducted remotely using maps and photos available on the Google Maps website; therefore the purpose of a stopped truck could potentially be erroneous. However, the validations are still believed to be accurate since secondary stops (such as gas stations) are easily identifiable from remotely sensed imagery. This validation process takes some manual effort, however,
access to remote images via aerial maps and a random sample help reduce the overall burden.

The validation using Google Maps assumes that the locations identified as secondary stop events will always be utilized in this capacity, but there is a possibility for primary stop events to occur at secondary based locations. This is expected to be a rare case in practice. For example, a primary stop at a fuel station will occur occasionally from a tanker truck, but observe a much larger number of trucks stopped at the site for an actual fuel refill (secondary stop event). Consider some rough estimates of 9,000 gallons for the tanker truck in the former case and 200 gallons for the semi-truck in the latter case. In such a scenario, this leads to a 1:45 ratio of primary stops to secondary stops at the site, suggesting that the assumption of secondary stop purposes is accurate in the majority of cases. As a result of this approach, the validations (and trips developed in the next chapter) always assume that fuel stations are used for secondary stop events. This has implications if an analyst is interested in studying trips based on fuel deliveries since they will not exist in the processed datasets discussed in this thesis.

An alternative to the manual validation is plausible if sample datasets are available with known secondary stops and known primary stops. The most applicable data in this regard would be travel diary information, which can be difficult (and expensive) to obtain.

The devised entropy technique is novel as it enables transportation modellers to classify the majority of primary and secondary truck stops. Its simplicity results in a small time commitment towards implementation with potentially large payoffs. The identification of stop events as primary or secondary is a necessary step to properly clean the data for use in future models of truck movements using passive GPS data. The cleaned
records can then be used as a key input for both macroscopic and microscopic transportation demand models. The latter tools are able to address policy questions related to the movement of goods by truck on topics such as trade, network resilience, and congestion. From a scientific viewpoint, this section of the thesis addresses an interesting spatial science problem within the realm of big data mining. The research question revolves around the application of entropy (a well-established concept in information science) to dig into a massive GPS dataset to identify the purpose of truck stops over space. The conducted analysis not only proves that the proposed approach works but also establishes a set of empirical characteristics to describe the nature of entropy associated with the explored clusters.

2.5.6 Additional Classification Methods

While the entropy approach discussed in the previous sections provided a method of establishing some of the stop events that are secondary, the validations indicate that false negatives occur in some cases. More specifically, since 29.1% of the stop events with an entropy below the threshold value of 2.75 were validated as secondary stops, they would result in an erroneous classification. As a result, other data is combined with the GPS data to identify the remaining secondary locations that were not recognized by the entropy method alone.

Additional secondary stops were identified using firm databases purchased from InfoCanada (Canadian firms) and DatabaseUSA (U.S. firms). More information on these datasets can be found in Section 2.6.2. The stop events are classified as secondary if they are located within 250 meters of a firm with an SIC industry of 5541, indicating a gas station or service stop.
In consultation with Transport Canada, the absence of firms near a stop may also indicate that its purpose is secondary. Several reasons for this occurrence include residential parking between driver shifts and roadside parking for rest when other suitable locations are not available. To determine if these stops can be classified as secondary, a random sample of 250 stops with no nearby firms was drawn. A validation of these stops using Google Maps revealed that 200 (80%) of them were likely to be secondary stops while 50 (20%) were identified as primary stops. Therefore the majority of the stops with no nearby firms are expected to be secondary. This is considered an acceptable error range in the case of this thesis since primary stops are used as end points for trips (as discussed in Chapter 3). Subsequently, stop events that have a primary purpose, but are erroneously classified as secondary, will result in fewer final trips. The overall philosophy adopted here is that a higher confidence in the observed trips is a fair trade off for a smaller dataset since the results still provide a very large amount of final trips.

While the classification introduced in this thesis differentiates between primary and secondary stops, the final information for each stop is still limited. For instance, we can be reasonably confident that goods are transferred at a primary stop, but the amount of goods and the direction of transfer (pick up or delivery) cannot be determined with any certainty. As a result, the trips generated from these primary stops as end points will have an unknown amount of goods. The trip may be transporting a full truck load (FTL) from the origin to destination, a partial less than truckload (LTL) shipment from the origin to destination, or an empty backhaul after a vehicle has off-loaded goods but not yet picked up more goods. Future research on the volume of goods hauled by the vehicle at any given time could be
conducted with the help of a fuel consumption variable that may be possible to obtain from some GPS data sources.

2.6 Stop Event Industry Identification

To determine the industry associated with a stop event, an analysis suggested that the simplest approach, the nearest firm, provides the best results. Moreover, a distance of 200 m was adopted as the maximum range to find the nearest firm, since larger distances decrease the likelihood of the stop event belonging with the given firm. The exception to this distance is the secondary firms, which use an expanded buffer of 250 meters. The larger distance for secondary firms is imposed due to the size of many secondary stop properties since they often accommodate a large volume of vehicles with similarly large parking lots.

A drawback to this approach arises from properties that are substantially larger than typical lots. In such a case, the distance between the stop event and the point representing the firm may exceed the chosen buffer size. This could be the cause for the low proportion of agriculture, mining, and forestry firms represented in the by GPS data (see Chapter 4 for details). But the alternative of larger search buffer distances opens up a greater likelihood that the stop event does not belong to the nearest identified firm.

Another issue arises from the possibility of businesses located closely within one building (such as a strip mall or plaza). From a data perspective, this results in multiple points stacked on top of each other to represent the different firms located at the same location. In such cases, the correct firm cannot be identified with 100% certainty based on the amount of information provided. However, the approach adopted in this thesis was the selection of industry based on frequencies observed in the dataset. For this purpose, a
A random sample of 1,000 stop events formed the basis of these frequencies where the primary industry was identified using Google Street View Maps. The 1,000 stop events were drawn in four increments with 250 draws each. Comparisons between the four sets found a strong consistency between each of them in terms of industry and location. The results from this random sample by industry are provided below in Table 2-2, along with the priority based on frequency. In the case of a tie-breaker scenario where multiple firms are located nearest to a stop event, the industry with the best priority (highest frequency) found at the location is selected. A possible downside to this approach is that low priority industries that were not observed as frequently will tend to be further neglected. Further information on industry bias is provided in Chapter 4.

Table 2-2: Validated industries from 1,000 randomly selected stop events

<table>
<thead>
<tr>
<th>SIC (2 Digit)</th>
<th>Description</th>
<th>Frequency</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-47</td>
<td>Transportation</td>
<td>324</td>
<td>1</td>
</tr>
<tr>
<td>52-59</td>
<td>Retail Trade</td>
<td>214</td>
<td>2</td>
</tr>
<tr>
<td>20-39</td>
<td>Manufacturing</td>
<td>165</td>
<td>3</td>
</tr>
<tr>
<td>15-17</td>
<td>Construction</td>
<td>58</td>
<td>4</td>
</tr>
<tr>
<td>70-89</td>
<td>Services</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>01-09</td>
<td>Agriculture, Forestry &amp; Fishing</td>
<td>20</td>
<td>6</td>
</tr>
<tr>
<td>50-51</td>
<td>Wholesale Trade</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>10-14</td>
<td>Mining</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>99</td>
<td>Unclassified</td>
<td>165</td>
<td>N/A</td>
</tr>
</tbody>
</table>
2.6.1 Firm Identification Tool

The identification of the nearest firm to a stop event is a simple endeavor, but can be time consuming if performed manually or with common GIS software tools. To this end, a firm identification tool was created to provide a streamlined method of inputting a set of point locations and returning the nearest firms and associated industries. For this thesis specifically, the point locations consist of stopped truck events identified in the GPS data where a truck remained relatively stationary for 15 minutes or longer, but the software is designed more generally to utilize any type of event with an associated longitude/latitude.

An image of the tool is provided in Figure 2-14 with locations loaded into the software for processing on the left (the locations displayed are not actual stop events to protect confidentiality). Data points can be loaded into the software manually for individual locations or from a separate file for batch processing. The user can choose to search for the nearest firms by drawing from online databases, a local database, or some combination of databases with each source colour coded for visualization. For the purpose of this thesis, the local database of firms includes a comprehensive set of 2013 firms located in Canada (InfoCanada) and the United States (DatabaseUSA).

A visual map of the input location and the nearest firms (denoted by the colour of the firm dataset and rank by distance 🏅) is shown on the right side of the software tool in Figure 2-14. Finally, the bottom of the software tool allows for field mapping to specific data fields from the input file and customization of the necessary parameters for the search such as the size of the search radius or the desired number of nearest firms.
2.6.2 Firm Datasets Tested

Several data sources were tested to identify a suitable firm dataset that can provide a comprehensive list of firms and industries. Each dataset is accessed in the software application using an application program interface (API) provided by the company, with the exception of data purchased and stored on a local server. Data from several companies that were tested include Google, Factual, and YellowPages.

The Google API provides similar results to a user search on the Google Maps website. The primary advantage to using the API instead of the public website is that the search can be performed for multiple points simultaneously. However, the data provided by the Google API has several disadvantages. First, a query limit of 500 per day (at least without adding costs) reduces the maximum processing capacity using this source. Secondly, the industry type is often unknown, resulting in a generic ‘establishment’ classification. Finally, the overall count of firms was found to be low, resulting in an increased likelihood of false positives (where the closest firm is not correct). An evaluation
of firms within the Windsor downtown area indicated that only 517 out of 3,370 downtown firms (16%) were found in the Google database.

The second tested source of data comes from Factual. The data is similar to Google but generally provides a more detailed description of the industry type by employing a tiered classification (general industry → intermediate industry → detailed industry). In addition, no query limit is imposed (unlike the Google API). However, the main drawback observed for the Factual data is that it is also not comprehensive. A comparison of Factual firms to a previously acquired dataset (InfoCanada) found that only 24% of the firms located in the City of Windsor were present in the former Factual dataset.

The YellowPages API was also tested but found to be lacking due to trial constraints and a limited query availability. For trial versions of the API, only 1 query is allowed per second (2 queries per second if a license is purchased) and requires a search for a specific industry (all industries therefore require over 20 individual queries for 1 location).

Complete datasets of firms from other sources were purchased by the Cross-Border Institute for this research in light of the issues encountered with the data sources described above. The U.S. data was purchased from DatabaseUSA and includes almost 14 million firms. Canadian data containing almost 1.4 million firms were purchased from InfoCanada. For comparison, the business pattern data published by Statistics Canada for December 2013 indicates that there are approximately 1.2 million businesses in Canada with a determinant number of employees (Statistics Canada, 2013). This number rises to 2.7 million when including businesses with indeterminate employment levels, but this increase is due to businesses that have a workforce with only contract workers, family members, or business owners. These employment-indeterminate businesses are therefore assumed to be
mostly self-employed entrepreneurs that would not be in the InfoCanada data and would also lack usefulness to this dissertation.

### 2.7 Data Processing Approach

To this point in the thesis, several approaches have been discussed, but it is useful to outline a full procedure of the GPS data processing used here. This processing begins by uploading the raw GPS data into an SQL database. Microsoft SQL Server 2008 R2 was chosen to match software utilized by counterparts at Transport Canada when the research first began. After this point, a number of scripts are executed to process the raw data into a more desirable form. The scripts are presented below in Figure 2-15 and discussed in the remaining paragraphs of this section to conclude the chapter.

![Figure 2-15: Microsoft SQL scripts for GPS processing](image)

Script 1 creates a series of empty tables with necessary fields that will be populated by data in future scripts. For example, a ‘stops’ table is created to store all of the stop events
that will be identified in the dataset. Script 2 maps the raw data into appropriate fields within a processed data table. A number of attributes are also created in this script including: a new ID for each vehicle that combines the carrier and power (truck) ID fields to create unique vehicle identifiers; a sequential ID for each GPS ping (after sorting the data by time for each truck); an elapsed time between subsequent pings for a given truck; the dwell time for a stopped truck; and a field to denote stop events. These stop events are identified using the approach discussed in Section 2.4.

The second script also includes the removal of intermediate GPS pings between the start and end of stop events. Consider an example with 10 pings occurring during a stop event. In such a case, the first and last pings are retained, but the 8 intermediate pings are removed since they do not provide any additional information. Performing this data cleaning removes roughly half the GPS pings in the processed dataset.

The second script concludes by indexing several numerical variables to vastly improve the processing performance when querying the dataset. These indexes provide a quick retrieval of data based on a given order (ascending or descending). For example, to find all pings within a given time period, the index provides the digital location of the appropriate pings to avoid searching the entire dataset and therefore speed up the processing. Variables that are indexed include the ping ID, truck ID, time stamp, latitude and longitude.

Script 3 provides a carrier summary that populates a table with the list of each carrier (by carrier ID), the number of trucks in the dataset for each carrier, and the number of pings in the dataset corresponding to each carrier. Besides providing basic information
on each carrier, the table later serves as a suitable source for scripts that need to process each carrier individually using looping iterations.

Script 4 utilizes the algorithm discussed in Section 2.3.2 to identify shipping depots for the individual trucking carriers. Script 5 populates a table for stop events based on the processed data from Script 2. Script 6 implements the entropy method discussed in Section 2.5.2 to determine the entropy (variety of carriers) of stop events surrounding each individual stop event. If a stop event exhibits an entropy exceeding 2.75, it is considered a secondary stop.

After Script 6 is implemented, the stop events are taken from the SQL server and processed through several additional software programs. As discussed in Section 2.5.5, entropy alone does not provide sufficient information to comprehensively classify all stop events. As such, the stop events are pulled from the SQL server and placed in the firm identification tool (see Section 2.6.1 for details) to assign the nearest firm (and industry) to each stop event. If a nearby firm is a gas station/rest stop (SIC code 5541), then it is classified as a secondary stop, even if there are closer firms by distance. In addition, stop events with no nearby firms are also classified as secondary stops (see Section 2.5.5 for details).

The data output from the firm identification tool is placed in a SAS workflow to process the resulting data automatically. Part of this SAS processing is a tiebreaker procedure when multiple firms are located at the minimum distance to a stop event. In such a case, industries that have been identified more frequently (based on a validation of 1000 stop events) are given precedence over other industries.
After the SAS procedure finalizes the industry for each stop event, they are transferred into ArcGIS software to determine the spatial zone (defined in Chapter 3) where each stop event belongs. Finally, the dataset of stops with a classification of stop purpose, industry, spatial zone is uploaded back into the SQL database. At this point, a table is created and populated exclusively with primary stop events and indexed for later queries.

Next, script 8 utilizes the primary stops as end points to determine inter-zonal truck trips travelling in Canada and the U.S. More information on this processing is provided next in Chapter 3 of this thesis. Moreover, Chapter 3 also contains information on the final SQL script (Script 9) in the processing workflow that identifies crossing times for truck trips at several major Canada-U.S. border crossing locations including the Ambassador Bridge and Blue Water Bridge.

2.8 Chapter 2 References


73
CHAPTER 3
TRIP CHARACTERISTICS

3.1 Introduction

The immense growth of satellite positioning technology over the last half century has led to an unprecedented amount of data that is generated and available for retrieval. One such technology, Global Positioning Systems (GPS), was first developed by the U.S. government in the 1970’s for military purposes. By 1995, GPS satellites had complete coverage over the globe. However, selective availability was employed by the U.S. government to degrade their GPS signals for civilian applications, resulting in a lower overall accuracy (Ghilani and Wolf, 2012).

Selective availability was removed at the turn of the century, enabling a faster spread of GPS technology for civilian uses such as navigation, surveying, and vehicle fleet tracking. The latter tracking application has been adopted extensively by large trucking carriers to trace their vehicle locations and optimize their routing and deliveries. This GPS data, when made available to freight transportation analysts, provides enormous opportunities to study the actual movements of vehicles and utilize the data as an input for transportation models. However, the GPS records are not originally created for such a purpose and must therefore be cleaned and processed into a viable format. The initial processing was the central focus of Chapter 2, including the identification and classification of stop events.
This chapter outlines a set of methods that utilize the GPS data to produce information pertaining to origin-destination trips for trucks that belong to Canadian owned carriers. Each data record in the raw GPS dataset corresponds to a ping where the location and time of a truck is recorded along with corresponding identification for the truck and carrier which is anonymous to protect their identities. The 2013 calendar year was adopted as the analysis time period, with over 1 billion initial data pings revealing the movements of approximately 56,000 trucks. The carriers belonging to this dataset tend to perform more long-haul trips compared to the average truck carrier (see Chapter 4 for details).

In addition to identifying trips from the GPS dataset, this chapter analyzes international trips that travel between Canada and the U.S. using the Ambassador Bridge or Blue Water Bridge. They were selected since they accommodate an astounding 28.9% and 14.2% of the value of goods transported by truck between Canada and the U.S. in 2013 (Bureau of Transportation Statistics, 2015). These goods are shipped by a large number of trucks crossing between Canada and the U.S. with truck vehicle counts of 2.4 million and 1.5 million for 2013 at the Ambassador and Blue Water, respectively (PBOA, 2015).

In the GPS dataset, 172,000 and 82,000 crossing events (for the 2013 calendar year) have been identified at the Ambassador Bridge and Blue Water Bridge. As such, the data represents 7.2% and 5.5% of all truck traffic across the two border crossings. Moreover, the industries pertaining to individual trips are estimated based on their start/end locations (primary stop events). The approach to estimate this industry was discussed previously in Section 2.6 of the thesis. As shown in Figure 3-1, these trips
provide information that combines the origin, destination, crossing time/location, and industry to capture the nature of international truck movements. To the author’s knowledge, this is the first study of its kind to provide detailed information about the characteristics of truck movements at these two major land crossings.

Figure 3-1: Principle attributes processed for each GPS derived trip

The remainder of this chapter begins by detailing the applications of GPS technology for transportation research. This is followed by an outline of the methods used to process the GPS data and identify trips. The calculation of border crossing events is then discussed, followed by results related to the characteristics of trips and crossing events before closing the chapter with concluding remarks.

3.2 Background

Transportation data has historically been derived from sample surveys of a population to determine their travel characteristics. This can include recall surveys such as face-to-face interviews, roadside surveys, and phone/internet surveys, as well as travel diaries that usually require a fairly active recording of events as the current day progresses. However, surveys can be expensive and labor intensive to conduct, while also relying on the respondent’s ability to correctly recall their movements (Stopher and
Greaves, 2007). Tracking the movement of an object using technology such as Global Positioning Systems (GPS) allows for an alternative method of data collection that tracks the actual movement of a passenger or vehicle. GPS data has been utilized in the recent past to study the travel movements of both passengers and freight vehicles. For passenger travel, GPS data can be used to identify the movements of people travelling between their home, work, and shopping activities (Wolf et al., 2001; Bohte and Maat, 2009; Shen and Stopher, 2013). Moreover, the data can be used as a means of identifying passenger trips across different modes of travel such as car, train, bike, and walking (Bohte and Maat, 2009; Xiao et al., 2015).

The availability of freight GPS data in the U.S. from the American Transportation Research Institute (ATRI) has led to a substantial presence of American studies utilizing GPS data for freight planning (Bernardin et al., 2015; Flaskou et al., 2015; Kuppam et al., 2014). However, the worldwide prevalence of GPS technology has also enabled its use in other locations such as South Africa (Joubert and Axhausen, 2011), United Kingdom (Hess et al., 2015), China (Yang et al., 2015), and Canada (Sharman and Roorda, 2013). The main advantage of GPS data is the large volume of information that can be used to observe the actual microscopic movement of trucks. This raw GPS data can then be converted into various useful forms such as truck routes (Hess et al., 2015; Ma et al., 2011), speed, and bottlenecks (Ma et al., 2011; Zhao et al., 2013), truck tours (Kuppam et al., 2014; Greaves and Figliozzi, 2008), and origin-destination trips (Bernardin et al., 2015; Ma et al., 2011; Zanjani et al., 2015).

Ma et al. (2011) derived origin-destination trips from GPS data between traffic analysis zones (TAZ) in Puget Sound, Washington. These trips were then placed in a
custom software interface to provide users with freight mobility measures between the origin and destination pairs. Bernardin et al. (2015) identified origin-destination trips for GPS data provided by ATRI for the U.S. states of Iowa and Tennessee. Their study emphasized the use of expansion factors to convert the origin-destination trips into an unbiased input for statewide freight models. Zanjani et al. (2015) also derived freight trips from GPS data supplied by ATRI, in this case for a statewide freight model pertaining to Florida. While the focus is on Florida, the study by Zanjani et al. produces origin-destination trips across the U.S. and Canada. The latter study indicates that GPS data is shifting toward a natural progression of tracking long-distance trips since the devices are not limited to a specific geographic location.

This thesis continues on this trend by focusing on international trips that cross the U.S.-Canada border at the Ambassador Bridge and Blue Water Bridge. Efforts have been made in the past to study delays and queueing events at border crossings using GPS data. For example, McCord et al. (2010) utilized onboard GPS devices to estimate crossing times at the Ambassador Bridge and Blue Water Bridge using detailed geofences enabling a separation of activities at the border such as inspections and duty-free visits. Their frequency of crossings for a roughly one year period included approximately 10,000 crossing events at the Ambassador and 5,000 crossing events at the Blue Water. McCord et al. (2014) expanded the previous study to examine queueing relationships at the two border crossing locations. In an earlier study, Goodchild et al. (2008) analyzed crossing times at the Pacific Highway crossing located in Blaine Washington using GPS data among other sources. Their total volume of GPS based trips included 44,000 crossing events from one carrier over the course of approximately three years (15,000 per
year). Moreover, their study utilized an additional survey to characterize the goods that typically travel across the Blaine border crossing.

By contrast, this thesis provides extra information regarding the trips observed with GPS data by determining their origin, destination, and industry. This information, along with the crossing time on the same trips, allows for immense opportunities to explore various trends that could help inform policy and advance discovery. Moreover, the large size of the GPS dataset in this thesis enables an exploration of the data over time with substantial sample sizes. For example, the resulting crossing events include 172,000 and 82,000 crossing events for the 2013 year at the Ambassador Bridge and Blue Water Bridge.

3.3 Methods of Analysis

3.3.1 Location and Type of Stops

The analyzed GPS dataset for this chapter includes the 2013 calendar year with over 1 billion GPS pings that are generated by the movement of 56,000 trucks owned by 850 Canadian carriers. Approximately 24.5 million stop events were observed when a truck is relatively stationary for 15 minutes or longer. The full set of stopped truck events was processed to provide more details such as the purpose for the stop. Primary stops occur when a transfer of goods takes place between the vehicle and location. Secondary stops occur to fulfill other needs such as fuel refills and driver breaks.

Several methods were devised to identify secondary stops in the dataset. This included an entropy measure quantifying the variety of carriers at a location, and proximity to firms denoted with industry codes identifying truck stop locations and gas
stations. In addition, the stop event is also identified as a secondary stop if there is no firm within a 200 meter radius of the stop location. Approximately 67% of the stop events were identified as secondary stops using the above approach. The remaining stop events were identified as primary stops. The industries pertaining to these primary stops were identified based on the nearest firm. More information on the processing approach was provided in Chapter 2 of the thesis.

The location of each primary stop was determined using a zoning system to identify the origin-destination (O-D) information for trips. On the Canadian side, this included census divisions delimited by Statistics Canada. On the U.S. side, the metropolitan statistical area (MSA) was used to define zonal boundaries. In addition, U.S. counties were utilized to fill in gaps in the U.S. since the MSA zones only exist in areas with significant development and/or urbanization. The combination of these three boundary types (census division, MSA, and county) resulted in 2,203 zones that reflect realistic political boundaries in lieu of a simpler rectangular boundary scheme. These zones are shown below in Figure 3-2.

![Figure 3-2: Zoning system applied to GPS based trips](image-url)
3.3.2 **Identifying Valid Trips**

A script was developed in Microsoft SQL Server, the platform used for data storage, to identify inter-zonal trips between origin and destination zones using primary stops as the trip ends (Script 8 in Section 2.7 of the thesis). The devised algorithm iterates through each origin-destination pair in turn. For a given zone pair, each primary stop event in the origin zone is evaluated to determine if a future primary stop event occurred in the destination zone. Initially, this may result in multiple potential trip records for only a single trip when multiple primary stop events occur in the origin zone. To remedy this issue, the potential trip with the later (more recent) stop event in the origin zone is retained while the other potential trips are discarded. This removes the actions of the truck during an intra-zonal activity which is not the focus of the thesis.

As an example, consider the scenario presented in Figure 3-3 to determine trips starting in Zone A and ending in Zone B from a vehicle that made 5 primary stops (P1 to P5). The initial algorithm processing results in two possible trips - P2 to P4 and P1 to P4. The algorithm then processes these trips to retain the P2 to P4 trip and discard the P1 to P4 trip as shown in Figure 3-4. As can be seen in this example, the primary stop events discussed in detail in Chapter 2 become important identifiers of trip ends in this approach. The intermediate primary stop P3 is not included as a trip end in this scenario since it is not located in Zone A or Zone B. However, P3 would be included as a trip end if the zone where it resides was selected as the origin or destination.
The outlined approach allows for the possibility of recording several individual trips from a single multi-leg trip, as shown in Figure 3-5. For example, an observed truck with primary stops in Toronto, Chicago, and Dallas may have 3 trips recorded: Toronto-Chicago; Chicago-Dallas; and Toronto-Dallas. This retains as much information as possible. However, caution is required when processing these trips later. For example, one single crossing event may exist for multiple trip records, therefore the duplication must be removed before tabulating crossing time statistics.
Figure 3-5: Example of a multi-zone trip

Up to this point in the process, some trips have been identified with very large time windows. To ensure that each trip is reasonably direct, a time limit constraint was applied to the resulting set of trips. Dispatcher data provided by Transport Canada for typical travel times between major Canadian and American cities was used as a baseline for this time restriction. As shown in Figure 3-6, a distance of roughly 900 km separates trips into two distinct groups based on effective speed. The effective speed includes both the travel time and any major break times for the driver. Short distance trips less than 900 km have a higher effective speed at approximately 70 km/hr since no large breaks are expected. The inclusion of substantial breaks for the longer trips noticeably lowers the effective speed to approximately 45 km/hr.
For each trip identified from the GPS data, the amount of time expected ($t_e$) for the truck to reach the destination can be calculated from the distance ($d_{ij}$) between the primary stops found in origin $i$ and destination $j$ as follows:

$$t_e = \begin{cases} \frac{d_{ij}}{70} & 0 \leq d_{ij} \leq 900 \text{ km} \\ \frac{d_{ij}}{45} & d_{ij} > 900 \text{ km} \end{cases}$$  

(Eq. 3-1)

where $t_e$ is the expected time that the trip should have taken based on the distance $d_{ij}$ between the trip ends in origin $i$ and destination $j$. A factor of 2 was used to account for abnormally congested travel and the underestimation of the true distance (since $d_{ij}$ is based on the Euclidean distance separating the origin and destination stops). This provides a maximum allowable travel time ($t_m$) for a given trip to be calculated as follows:

$$t_m = 2t_e$$  

(Eq. 3-2)

where $t_m$ is the maximum allowable trip time between a given pair of origin and destination stops to be considered valid, and $t_e$ is the expected travel time.
3.3.3 Border Crossing Time

In addition to the origin, destination, and industry of a trip, border crossing information was also estimated. The estimation of crossing time at the Ambassador Bridge and Blue Water Bridge were completed based on a geofence method that has been used by Transport Canada in earlier pilot projects. As shown in Figure 3-7, this included: an inner fence defining the area where the crossing time is estimated; a roughly 20 km by 20 km outer boundary to define the entry/exit jurisdiction (i.e. Canada or the U.S.); and a linear interpolation of the time between pings to estimate when the vehicle crossed the entry and exit thresholds. For a trip to be included as a valid crossing, there needs to be at least one ping in the inside geofence as well as pings in the outer boundary on both sides of the border. The size of the outer boundary can be adjusted based on the required circumstances, where a larger boundary will increase the number of valid observed events but also increase the potential variation for the interpolation of time between the outside pings and the entry/exit into the geofence zone.

The crossing times obtained from these geofences provide the total time it takes a truck to cross the area encompassing major border activities, but does not break this time down into specific functions. For example, time spent in the geofence will include time waiting on the bridge or plazas due to congestion, stopping at a duty free store, or being inspected at primary or secondary booths. Smaller geofences could be used to quantify these separate functions, but were not implemented with this GPS dataset due to the relative sparsity of ping intervals (discussed in Chapter 2). As a general guideline, smaller geofence zones will require smaller ping intervals to be successful.
After concluding this data processing, a dataset of origin-destination trips from the GPS data is developed. The industry associated with the trip ends and crossing time for international trips is combined to provide additional information. The results of these trips are provided in the next section.

3.4 Discussion and Results

Using the geofence technique in the previous section (Section 3.3.3), there were 172,000 observed crossing events that utilized the Ambassador Bridge and 82,000 observed crossing events that utilized the Blue Water Bridge (for the 2013 calendar year). However, not all observed crossing events have a corresponding processed trip (from Section 3.3.2) since gaps in the GPS data occasionally occur. As such, a total of 119,231 crossing events at the Ambassador Bridge had corresponding trips (representing 69% of
observed crossings). Similarly, 53,224 trips at the Blue Water Bridge had corresponding trips (representing 65% of observed crossings).

Useful statistics on the characteristics of the derived international trips are generated to characterize the nature of the truck movements occurring between Canada and the U.S. The statistics in Table 3-1 provide the total number of stops and the primary stops for a given month of the year along with the number of trips crossing the Ambassador Bridge and Blue Water Bridge. Correlations of 67% and 91% occur between the number of primary stops and the frequency of trips at the Ambassador and Blue Water crossings, respectively. This suggests that the proportion of international trips to the size of the dataset remains fairly consistent over time. This proportion of trips to primary stops ranges from 0.5% to 2%.

Table 3-1: Summary of data and trips by month (2013)

<table>
<thead>
<tr>
<th>Month</th>
<th>Total Stops (Sₜ)</th>
<th>Primary Stops (Sₚ)</th>
<th>Ambassador Trips (Ta)</th>
<th>Ta/Sₚ (%)</th>
<th>Blue Water Trips (Tb)</th>
<th>Tb/Sₚ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>2,757,370</td>
<td>914,345</td>
<td>11,522</td>
<td>1.26</td>
<td>5,872</td>
<td>0.64</td>
</tr>
<tr>
<td>February</td>
<td>2,437,177</td>
<td>794,132</td>
<td>10,477</td>
<td>1.32</td>
<td>4,913</td>
<td>0.62</td>
</tr>
<tr>
<td>March</td>
<td>2,736,241</td>
<td>933,274</td>
<td>11,056</td>
<td>1.18</td>
<td>6,015</td>
<td>0.64</td>
</tr>
<tr>
<td>April</td>
<td>2,612,681</td>
<td>866,424</td>
<td>11,124</td>
<td>1.28</td>
<td>5,716</td>
<td>0.66</td>
</tr>
<tr>
<td>May</td>
<td>1,470,577</td>
<td>501,992</td>
<td>5,549</td>
<td>1.11</td>
<td>3,333</td>
<td>0.66</td>
</tr>
<tr>
<td>June</td>
<td>1,957,234</td>
<td>708,080</td>
<td>10,827</td>
<td>1.53</td>
<td>4,480</td>
<td>0.63</td>
</tr>
<tr>
<td>July</td>
<td>2,030,949</td>
<td>731,082</td>
<td>9,898</td>
<td>1.35</td>
<td>3,906</td>
<td>0.53</td>
</tr>
<tr>
<td>August</td>
<td>1,858,422</td>
<td>670,779</td>
<td>10,428</td>
<td>1.55</td>
<td>4,367</td>
<td>0.65</td>
</tr>
<tr>
<td>September</td>
<td>1,681,476</td>
<td>606,067</td>
<td>11,887</td>
<td>1.96</td>
<td>4,455</td>
<td>0.74</td>
</tr>
<tr>
<td>October</td>
<td>1,752,001</td>
<td>636,732</td>
<td>11,140</td>
<td>1.75</td>
<td>4,406</td>
<td>0.69</td>
</tr>
<tr>
<td>November</td>
<td>1,496,340</td>
<td>531,291</td>
<td>7,576</td>
<td>1.43</td>
<td>2,769</td>
<td>0.52</td>
</tr>
<tr>
<td>December</td>
<td>1,714,151</td>
<td>611,139</td>
<td>7,747</td>
<td>1.27</td>
<td>2,992</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Correlations: rₛₚ,ₜₐ = 67%; rₛₚ,ₜₚ = 91%
The results from this dataset suggest that 52% of the distance travelled pertains to trips that cross the Canada-U.S. border. This is substantially higher than 14% of the overall distance travelled by all Canadian truck trips based on 2009 data available from Statistics Canada (based on data from Statistics Canada – Table 9-2, 2009). This confirms suspicions that the GPS data is heavily biased towards long-haul trips that occur between Canada and the U.S. A bias towards longer distance trips is discussed in Chapter 4.

3.4.1 Spatial Distribution of Trips

Trip productions by zone is shown in Figures 3-8 to 3-11 for Canada and the U.S. trips that use the Ambassador Bridge and Blue Water Bridge. Trip attractions are not included for brevity since the observed patterns are very similar. The maps were created using natural breaks in the symbology to differentiate between zonal frequencies within each map. As a result, comparisons between figures should be made with caution given the variability between color schemes. The figures represent the spatial distribution of trips for a full year. However, results show similar results throughout the calendar year. A correlation analysis was conducted to further establish the consistency over time by comparing the distribution of trips between two consecutive month pairings (i.e. January and February, February and March, etc). All of the month to month correlations are 95.3% or higher, suggesting a similar spatial distribution of trips over time.

From a Canadian perspective, the trips in the GPS dataset that utilize the Ambassador Bridge are concentrated along the Highway 401 corridor between Montreal and Windsor with the heaviest activity occurring in the Region of Peel in the Greater Toronto Area. A similar trend occurs for the Blue Water Bridge in Figure 3-9 with the
exception that trips are shifted in southern Ontario along Highway 403 to Sarnia. Their similarities are expected based on the close proximity of the crossings. For example, trips between Toronto and Chicago take approximately the same amount of time to complete when crossing the Ambassador Bridge and Blue Water Bridge (this specific case is expanded upon in Chapter 5). Figures 3-8 and 3-9 also display the production of trips from cities residing in the western provinces (such as Vancouver, Edmonton, and Calgary) and eastern provinces (such as Quebec City, Moncton, and Halifax).

Figure 3-8: Origin trips using the Ambassador Bridge by census division (2013)
Figures 3-10 and 3-11 show the pattern of trip production on the U.S. side of the border. The highest concentration occurs for cities closer to the border such as Chicago, St. Louis, and Cincinnati. However, a large number of distant trips are gravitating from the southern edges of the continental U.S. in Laredo and Los Angeles. The Laredo based trips presumably represent a connection to Mexico while the Los Angeles shipments likely represent Canadian trade across the Pacific Ocean.

An interesting phenomenon can be seen in the resulting maps where Canadian zones exhibit larger amounts of clustering compared to the U.S. maps. We have confirmed the larger presence of clustering in Canadian zones using a global Moran’s I.
(MI) statistic using the GeoDa software package. The MI statistic for Canadian trip production is 0.382 (p < 0.001) and 0.246 (p < 0.001) for trips crossing the Ambassador Bridge and Blue Water Bridge. For U.S. trip production, the corresponding MI values are 0.056 (p < 0.006) and 0.034 (p < 0.008). These results suggest that the U.S. spatial patterns are more spatially dispersed compared to Canadian zones. However, clustering is still noticeable in the U.S. with 99% confidence.

Figure 3-10: Origin trips in US zones through the Ambassador Bridge (2013)

Figure 3-11: Origin trips in US zones through the Blue Water Bridge (2013)
3.4.2 Density Maps of International Trip Productions

For a contrasting method of visualization, heat maps derived from kernel density estimations are provided below in Figures 3-12 to 3-17. These maps parallel the figures provided in the previous section, but also include an extra year of information (comparisons of 2013 and 2014 revealed very similar patterns) and the addition of maps pertaining to the Peace Bridge. The latter is a high volume border crossing in Ontario that connects Fort Erie, ON with Buffalo, NY. Other crossings could be easily added in the future by creating geofences at the appropriate locations to determine crossing events and connecting them to the trips in the GPS dataset.

The heat maps of trip productions confirm the relative similarity of spatial patterns between the Ambassador Bridge and Blue Water Bridge. However, the Peace Bridge reveals a very different pattern. On the Canadian side, the Peace Bridge generally services local trips that originate in the Greater Toronto Area (GTA) and surrounding regions while Montreal is a much smaller source of trips. This is likely due to the Ambassador/Blue Water Bridges taking most trips from Montreal that are headed west or south while other crossings exist in Quebec to allow trucks to travel to the east. On the U.S. side, the density map reveals that the Peace Bridge tends to facilitate trips that start from eastern states.
Figure 3-12: Canadian density of truck trips crossing the Ambassador Bridge

Figure 3-13: U.S. density of truck trips crossing the Ambassador Bridge
Figure 3-14: Canadian density of truck trips crossing the Blue Water Bridge

Figure 3-15: U.S. density of truck trips crossing the Blue Water Bridge
Figure 3-16: Canadian density of truck trips crossing the Peace Bridge

Figure 3-17: U.S. density of truck trips crossing the Peace Bridge
3.4.3 Border Crossing Time

The estimated crossing time information was aggregated to produce crossing time distributions for the two bridges as shown in Figure 3-18. The patterns that emerge in these figures are typical of border crossing times. That is, the distribution is substantially skewed in which the average crossing time occurs at a relatively low value followed by a heavy right tail. The extreme values on the right end of the distribution reflect the occasionally large delays leading to significant 90th and 95th percentile crossing times that potentially add substantial costs to international trips. More information on the negative impact of variability at the border on supply chains can be found in Anderson and Coates (2010).

![Figure 3-18: Border crossing time distributions (2013)](image)

Industry specific crossing times for the processed trips are presented in Table 3-2. The average crossing time by industry at the Ambassador Bridge varies between 21.2 – 23.9 minutes for trucks crossing into the U.S. and 16.5 – 18.0 minutes for trucks crossing into Canada (excluding sectors with low trip counts). However, since the variability in
the distribution of crossing time is potentially more important, values pertaining to the 95th percentile were also analyzed. The 95th percentile values can highlight unexpected delays, which could lead to a complete disruption of production lines for Just-in-Time (JIT) deliveries. At the Ambassador Bridge, a range of 10 minutes can be observed for the 95th percentile crossing time with values ranging from 52.4 – 62.9 minutes and 36.0 – 46.2 minutes for U.S. bound and Canada bound traffic.

At the Blue Water Bridge, the variations by industry are slightly higher. The average crossing time ranges from 14.9 – 21.3 minutes and 13.7 – 19.1 minutes for trucks crossing into the U.S. and Canada, respectively. The 95th percentile values at the Blue Water Bridge have similarly larger ranges compared to the Ambassador with values from 37.7 – 60.8 minutes and 36.2 – 50.4 minutes.

The crossing time at the Ambassador Bridge is slightly higher compared to the Blue Water Bridge by a few minutes on average. This difference may be attributed to the size of the geo-fences drawn around the actual border crossing plazas, which are dependent on the physical infrastructure layout/constraints. Another explanation is that the Ambassador Bridge experiences higher volumes of trucks. In addition, this chapter does not go into details regarding the influence of temporal variables such as time of day on the crossing choice. However, an analysis on this subject can be found later in Chapter 5. Finally, the choice of crossing can be influenced by different pricing schemes since most crossings charge a flat fee per axle while the Ambassador Bridge includes adjustments based on weight.
## Table 3-2: Crossing times by aggregate destination industry

<table>
<thead>
<tr>
<th>SIC Aggregate Industry</th>
<th>Total Trips</th>
<th>To USA</th>
<th>To Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Avg. Crossing</td>
<td>90th Percentile (Min)</td>
</tr>
<tr>
<td>Agriculture, Forestry &amp; Fishing</td>
<td>633</td>
<td>402</td>
<td>23.37</td>
</tr>
<tr>
<td>Mining</td>
<td>103</td>
<td>94</td>
<td>25.10</td>
</tr>
<tr>
<td>Construction</td>
<td>4,057</td>
<td>1,809</td>
<td>23.78</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>15,874</td>
<td>8,828</td>
<td>21.23</td>
</tr>
<tr>
<td>Transportation</td>
<td>21,167</td>
<td>8,633</td>
<td>23.78</td>
</tr>
<tr>
<td>Communications &amp; Utilities</td>
<td>262</td>
<td>185</td>
<td>23.49</td>
</tr>
<tr>
<td>Wholesale Trade</td>
<td>8,497</td>
<td>5,684</td>
<td>22.49</td>
</tr>
<tr>
<td>Retail Trade</td>
<td>14,464</td>
<td>9,105</td>
<td>22.82</td>
</tr>
<tr>
<td>Finance, Insurance &amp; Real Services</td>
<td>1,234</td>
<td>944</td>
<td>23.38</td>
</tr>
<tr>
<td>Total Ambassador Trips</td>
<td>51,434</td>
<td>22,413</td>
<td>18.79</td>
</tr>
</tbody>
</table>

### Blue Water Bridge

<table>
<thead>
<tr>
<th>SIC Aggregate Industry</th>
<th>Total Trips</th>
<th>To USA</th>
<th>To Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Avg. Crossing</td>
<td>90th Percentile (Min)</td>
</tr>
<tr>
<td>Agriculture, Forestry &amp; Fishing</td>
<td>224</td>
<td>118</td>
<td>24.42</td>
</tr>
<tr>
<td>Mining</td>
<td>64</td>
<td>39</td>
<td>26.14</td>
</tr>
<tr>
<td>Construction</td>
<td>2,589</td>
<td>747</td>
<td>20.21</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>8,573</td>
<td>3,167</td>
<td>18.94</td>
</tr>
<tr>
<td>Transportation</td>
<td>9,464</td>
<td>3,124</td>
<td>19.59</td>
</tr>
<tr>
<td>Communications &amp; Utilities</td>
<td>139</td>
<td>87</td>
<td>22.59</td>
</tr>
<tr>
<td>Wholesale Trade</td>
<td>5,851</td>
<td>4,379</td>
<td>14.92</td>
</tr>
<tr>
<td>Retail Trade</td>
<td>5,642</td>
<td>3,193</td>
<td>20.68</td>
</tr>
<tr>
<td>Finance, Insurance &amp; Real Services</td>
<td>369</td>
<td>233</td>
<td>21.26</td>
</tr>
<tr>
<td>Services</td>
<td>6,445</td>
<td>2,583</td>
<td>20.73</td>
</tr>
<tr>
<td>Public Administration</td>
<td>329</td>
<td>200</td>
<td>19.03</td>
</tr>
<tr>
<td>Non-classifiable Establishments</td>
<td>1,041</td>
<td>447</td>
<td>18.66</td>
</tr>
<tr>
<td>Total Blue Water Trips</td>
<td>40,730</td>
<td>18,317</td>
<td>18.79</td>
</tr>
</tbody>
</table>

Categories with low sample sizes are shown in grey to emphasize a lower confidence in their output.
3.4.4 Effect of Distance on Crossing Times

The average crossing time was sorted for trips into distance bins of 100 km intervals as shown in Figure 3-19. Trips with a distance greater than 2,600 kilometers were censored and aggregated together due to their low frequency. Several notable characteristics can be seen in the resulting graph. First, international trips with distances below 100 km had noticeably smaller crossing times. Secondly, the average crossing time tends to rise as the trip distance increases. As such, trend lines were fitted to the graph using a power relationship.

Several possible explanations exist for the exhibited trend. Obviously, short trips that start and end near the border are in a better position to adjust schedules to avoid large delays at the border. Moreover, firms located near the border are more likely to be familiar with temporal variations in cross-border traffic. Such patterns could be investigated further in this data to confirm the presence of this advantage. Finally, the crossing time at the border will have a larger impact on short haul trips since the distance is smaller. In such a case, shorter trips become more sensitive to longer wait times at the border. For example, a 25 minute crossing may increase a short trip from 30 to 55 minutes while a longer trip may increase from 500 minutes to 525 minutes. The former scenario results in an 83% increase in time while the long trip only observes a 5% increase. The proportion of a trip that is comprised of delay (expected and unexpected) is a central theme in Chapter 6 of the thesis.
3.5 Conclusions

This chapter outlines the methods and results arising from a conversion of freight GPS data into observable trips for Canadian trucks traveling between Canada and the U.S. The final outcome of this process is a dataset of trips utilizing the border crossings at the Ambassador Bridge and Blue Water Bridge with information pertaining to the origin and destination of the trip along with industries based on the trip end points and the corresponding border crossing time. This is a timely subject given the increasing availability of large passive datasets and an interest in converting GPS data into useful inputs for transportation models and policy analysis.

The methods of deriving trips from GPS data provide analysts with useful guidance to apply the information to their own work. This makes these methods practice ready given the strong similarities across most GPS datasets. The largest difference in
methods applied to other GPS datasets will likely be the treatment of stop identification as our dwell time of 15 minutes was selected based on the long distance preference of our carriers. Other datasets may subsequently require a shorter dwell time threshold when short urban deliveries are more prominent. However, the chosen cut-off value needs to be handled carefully as shorter values can increase the likelihood of false positive stops arising from congested traffic conditions.

While a number of past studies have identified trips from passive GPS data, this analysis derives international truck trips with attributes on crossing time and industry type from such data. These key trip characteristics are valuable for future research since the derived data can be dissected in multiple ways. As an example, results on the crossing time information for two border crossings by industry and trip distance were discussed. From these results, it is found that short trips tend to have noticeably shorter crossing times at the border for both the Ambassador Bridge and Blue Water Bridge.

The results of this chapter show a strong consistency in the spatial patterns and nature of the international trips from month to month. This suggests that the carriers in the dataset typically follow similar patterns over time to service their regular clients. As a result, the data processed from GPS pings can be confidently utilized for models in the present and short-term future to benefit from the additional knowledge gained from the characteristics of vehicle behavior.

The methods described in this chapter will be particularly useful for transportation researchers given the standardization of information obtained from GPS data. Moreover, the results provide a strong foundation for freight transportation/trade models between Canada and the U.S., although the data here only utilizes Canadian trucks. The obtained
results also produce valuable insight on crossing time distributions by type of industry. Such information can be used in the development of predictive economic trade models. Besides modelling, the industry information for these trips can also be useful for policy analysts when combined with crossing data and the trip end points.

3.6 Chapter 3 References


4.1 Introduction

A prominent method of data collection for passenger and freight transportation is the application of Global Positioning System (GPS) devices to capture vehicle movements. This trend is the result of a successful integration of GPS technology as a commonplace occurrence for navigation. Moreover, commercial firms employing fleet vehicles for goods movement widely adopted GPS to remotely observe their vehicles and deploy their resources accordingly. The data generated from these GPS devices result in data pings - individual spatiotemporal points that identify where a vehicle was located over space at a particular instance of time. These pings also typically denote a particular vehicle and company using some form of identification, though this is often anonymized to protect the identity of the drivers and firms. Using the identification information, GPS pings for a particular truck can be combined together to observe the movements of the vehicle over time and convert this information into vehicle trips.

The large output of data generated from GPS devices, and the increasing availability of such information from vendors such as the American Transportation Research Institute (ATRI), are positive features that position the GPS data source as a viable alternative to traditional data collection models. Moreover, the data generated from GPS pings avoid potential recall errors that may occur from surveys requiring a respondent to reconstruct past activities (Stopher and Greaves, 2007). However, this data source also
carries some potentially negative drawbacks. The data does not typically provide explicit information on the nature of activities carried out by trucks. This is common for GPS data derived from commercial fleets with no expectation of their data adapted later for transportation models. By comparison, there are examples of data surveys that are based on travel diaries to supplement the GPS data with activity information (Du and Aultman-Hall, 2007). Such surveys are typically designed from the ground up for modelling purposes and require a higher level of involvement from the survey respondent. In turn, this can lead to additional compensation to the respondent, thereby increasing the cost of the survey and typically limiting it to a lower number of participants.

Post-processing methods on passive GPS datasets obtained from fleet tracking companies can provide an alternative to the high costs associated with devising specialized surveys. For example, Chapter 2 of this thesis employed entropy as a method of differentiating the purpose of stops as primary (to transfer goods) or secondary (for truck driver/vehicle needs). Sharman and Roorda (2011) tested various clustering techniques to group trip ends together when they occur at the same facility. As another example, Bohte and Maat (2009) used vehicle speeds and spatial proximity to geographic features to determine the trip purpose and mode of transportation for passenger trips observed from GPS data.

These emerging efforts to process GPS data are promising for the transportation field since they can be used to analyze the patterns inherent in the data and utilize other spatial information to infer the activity patterns. As such, new algorithms to process Big Data are an integral component necessary to provide a viable alternative to traditional data sources. However, the application of different methods on such data has to consider the
potential bias resulting from a non-representative sample. Failure to address this concern can lead to erroneous conclusions about the movement patterns.

Bricka et al. (2009) discovered differences among the demographic characteristics of responders for traditional surveys and GPS based surveys used to observe household travel patterns. For commercial vehicles, a GPS dataset often comes from a single GPS service provider that is employed by one or more commercial companies to supply the devices and software necessary to track a fleet of vehicles. The set of companies that use a single service provider are more likely to be interested in similar services since each service provider may provide differentiated products. Moreover, the characteristics of the firms that require a GPS service provider may differ from other firms that do not require such extensive tracking services. For example, trucks with local routes or consistent schedules will not necessarily require the resources of a GPS service provider.

This chapter discusses the issue of representation in the GPS truck data and offers a procedure for expanding a sample of truck trips travelling between census divisions in Ontario. The contribution of this work is three-fold. First, the biases in the GPS data are important to consider when viewing many of the results presented in this thesis. These biases are also likely present to some degree in other studies that utilize passive GPS data as a source of information. Secondly, the expansion method provides a means of increasing the sample data to match observed trip totals. Finally, the resulting expansion factors provide trip rates per firm by industry for aggregate freight trip generation in the Canadian context. However, these trip rates are fairly general in nature, therefore microscopic models would likely require more complex equations for trip generation.
The rest of this chapter begins by discussing some past efforts in literature to perform trip generation and expand transportation data samples. The chapter then describes the GPS dataset and the biases encountered in the data. Next, a novel approach is provided to reduce the issue of bias in the dataset and expand the data sample to match aggregate totals. The methods and validation results of this study on the Province of Ontario are provided along with a final discussion.

4.2 Background

4.2.1 Freight Trip Generation

In most travel demand models, the calculation of the number of trips produced and attracted to an individual location or zone is an early task. Notably, the four-stage urban transportation planning system (UTPS) model approach includes trip generation as the first step (Ortuzar and Willunsen, 2011). In freight transportation, the type of trip generation employed by a modeller is often dictated by the type of intended model. For example, NCHRP report 298 (TRB, 2001) describes two approaches – (1) truck based models and (2) commodity based models. The latter type of models often utilize a payload conversion factor to transfer commodities by tonnage or value into some quantity of trucks. However, the truck based models are more relevant to this thesis as GPS data tracks individual vehicles.

The commonly cited method of trip generation for truck based models are trip rates based on a variable consistent with land use categories such as square footage by industry (TRB, 2001). The most frequently used variable in this regard is employment, which leads to a trip rate based on the number of total employees by industry for a given location or zone. For example, a table provided in the Quick Response Freight Manual II (Cambridge
Systematics – Table 4.1, 2007), based on the Phoenix Metropolitan Urban Truck Model, provides a sample of truck trip rates by industry employment (and households).

However, Holguín-Veras et al. (2011) contend that employment is better suited towards freight generation (as a term for the volume/value of economic goods produced) instead of freight trip generation. The authors assert that this is due to the non-linear correlation between the volume of goods produced and the number of trips due to factors such as shipment size and frequency that can be adjusted to optimize costs. Their study then compares three types of trip generation models for individual firms by industry including (1) trip rates by employment, (2) trip rates by firm (constant with no variation by employment), or a hybrid of the two represented as a small linear regression based on employment rate (the slope) and by firm (the constant). Their results indicate that employment based trip rates are the best option in only 18% of cases (by industry) while using only the constant (trip rates by firm) is the best option in 52% of cases. As will be shown in this chapter, the latter trip rate by firm is the approach utilized here. This is partially done out of necessity, however, since it is believed that the employment levels that pertain to the firm data used in this thesis are not accurate enough for this task.

Three main data sources for truck trip generation are traditionally used including direct counts of trips, roadside intercept surveys and travel diary surveys (TRB, 2001). However, GPS data could provide an alternative source of information for trip generation. For example, the Oregon Department of Transportation has considered using their cellphone based GPS data - Truck Road Use Electronics (TRUE) - to calculate trip generation rates by vehicle type (Bell and Figliozzi, 2013). As another example, truck tours are now often modelled with GPS data, such as the tour model developed by
Doustmohammadi et al. (2016). The authors used GPS data and a calibrated model for tour
generation that included variables such as employment, population, land area, and zone
type.

4.2.2 Sample Expansion

The expansion of a data sample is a common occurrence in numerous disciplines
to produce a dataset from the sample that matches the full population/universe with respect
to its size and desired characteristics. One of the most well-known examples is census data,
where detailed information is not normally collected for each individual of a nation. For
instance, the Canadian census includes two sections – (1) a short form with basic
information filled out by the entire population and (2) a long form that contains more
detailed information but only made mandatory for a sample of 20% of the population which
includes every 5th house (aside from the 2011 census where the survey was voluntary).
Weights accompany the (confidential) long form survey results to provide an unbiased
representation of the entire population from the sample (see Roberts, 2015, for more
details).

The transportation field also relies on sample expansions since surveys and data
collection can be costly and infeasible for an entire group/population. For example, 2014
Transportation Panel Survey (CH2M Hill, 2015) conducted for Vancouver BC used
weighting/expansion to adjust the 0.48% sample of survey participants to match the full
population. The sample was expanded based on demographic characteristics such as age
and gender. Milligan et al. (2016) discuss the common practices in transportation of
expanding short term traffic counts into an Annual Average Daily Traffic (AADT) volume.
Their paper utilizes an existing individual permanent count (IPC) method that expands a short-term count station with temporal results from a suitable permanent control station.

4.3 Primary Data and Biases

The GPS dataset originally obtained from Shaw Tracking (via Transport Canada) forms the basis of analysis in this study. A subset of the data pertaining to the month of January, 2013 was utilized. The January data pertains to approximately 730 Canadian owned trucking firms and 40,000 individual trucks. Processing the GPS dataset produced 250,000 trips representing trucks that travel within Canada and across the U.S. These trips were derived from inter-zonal truck movements between census divisions (in Canada) and metropolitan statistical areas / counties (in the U.S.).

The initial processing of the raw GPS data was discussed in Chapter 2. This included the identification of vehicle stop events and classification of these stops as primary or secondary. A detailed description of the processing used to derive these trips can be found in Chapter 3 of the thesis. The latter chapter described the creation of a dataset of trips bounded between primary stop events, and the development of a time based constraint to determine an allowable travel time for a reasonable trip. In addition, the industry of the trip was estimated based on the nearest firm (within 200 meters) to the stop location. The list of approximately 507,660 Ontario firms, purchased from InfoCanada, included their location and industry. Since each trip is bounded by a primary stop at each end of the trip, an origin industry and destination industry are both estimated.
4.3.1 Distance Bias

An analysis of the GPS data led to the discovery of several biases inherent in the sample of trips. For example, a comparison was performed between the GPS derived trips (in Chapter 3) and the 2006 commercial vehicle survey (CVS) data created by the Ministry of Transportation Ontario (MTO). The CVS origin-destination information is itself considered a biased source of information favoring inter-zonal trips over shorter urban trips. However, both datasets are used in this case to compare inter-zonal trips only. In addition, both datasets utilize the Ontario census division delineations as the zoning system. Outside of Ontario, the MTO dataset utilizes larger zones by aggregating to the province/state level. A distribution of the trips from each dataset was created based on the distances between zones as shown in Figure 4-1 (bins of 400 km were used). The data was standardized for comparison purposes by calculating the total proportion of trips for a given distance range.

In both cases, the frequency is highest for short range trips and reduces with an increasing distance. In fact, the MTO CVS dataset provides a fairly smooth curve that would fit well with negative exponential or power curves often associated with gravity models of trip distribution. By comparison, the distribution for the GPS dataset shows a lower proportion of short distance trips, while longer trips exhibit a higher overall proportion beginning with trips traversing more than 800 km in length. The higher proportion of longer distance trips matches expectations discussed in the introduction of the chapter since trucks/firms with short range trips are less likely to rely on a GPS service provider. Moreover, many of the shipping carriers in the dataset are large, for-hire companies that specialize in longer-distance freight deliveries.
4.3.2 Industry Bias

Industry bias is another area of concern with the GPS dataset. In Ontario, trips derived from the GPS data only utilize 5 different mining firms. By contrast, the firm dataset for Ontario purchased from InfoCanada estimates a total of 988 mining firms that exist in Ontario. Our sample therefore covers only 0.5% of these firms. By comparison, all categories of industry are represented by 9,097 firms in the GPS dataset and 507,660 firms in the InfoCanada firm database for a sample proportion of 1.8%. The proportional representation of firms by industry for the sample of GPS derived trips is provided in Figure 4-2. Manufacturing and Transportation exhibit higher proportions of representation, while primary industries (‘mining’ and ‘agriculture, forestry and fishing’) and services contain a lower proportional representation. For the service industry, the displayed result may be intuitive since only a portion of them require shipments by commercial trucks. However, a larger representation is expected for primary industries where goods distribution is more common.

Figure 4-1: Distribution of trips grouped by distance
The lower representation of primary industries may be the result of the method used to estimate the industry of a trip end. Several methods were attempted to provide the best results in this regard including the industry from the nearest firm and the most frequent industry in the nearby vicinity. Since the best method resulting from this testing was the nearest firm, a point to point relationship was established between each of the stop event locations and the nearest firm within 200 meters. However, the point location of a firm is sometimes located at the road entrance to the property since the address of the business may be used for geocoding. If a property is extremely large (as in the case of many primary industries where large land space is required), the actual point for the firm may be located outside the search radius and remain undetected. Utilizing lot boundary information is a potential method of mitigating this issue, but can be difficult to obtain. This is particularly true when observing this GPS dataset which covers both Canada and the U.S., where individual lots with business information would need to be obtained from each municipality independently.
4.4 Expansion Methods and Results

Based on the analysis described above, two major types of bias were identified from the GPS based trips: (1) a spatial bias where our dataset over-represented longer distance trips and (2) an industry bias where primary industries are particularly under-represented. This section describes the methods that were devised to reduce this bias while also expanding the sample data to match aggregate totals as shown in Figure 4-3. The numbers in Figure 4-3 represent the order of each step, and are used as a reference for the remainder of this section of the thesis.

Figure 4-3: Flow chart outlining trip expansion
4.4.1 Step 1 - Trip Rates

To begin, average trip rates per firm are estimated over a one month period and calculated as:

\[ R = \frac{T}{F} \]  
(Eq. 4-1)

where \( R \) is the trip rate, \( T \) is the number of trips (derived from the GPS data as discussed in Chapter 3), and \( F \) is the number of firms visited by those trips (derived from an InfoCanada firm dataset). Average trip rates for Ontario, the rest of Canada, and the U.S, are provided in Table 4-1 for trip productions and attractions. The results show that the trip rates for Ontario are the highest. The rest of Canada exhibits a slightly lower rate compared to Ontario while the U.S exhibits a substantially reduced trip rate. The large drop in trip rate for U.S firms is likely caused by the nature of the GPS data source tracking only Canadian owned carriers. As a result, U.S firms are visited less frequently.

Table 4-1: Average trip rates by jurisdiction

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Production</th>
<th></th>
<th></th>
<th>Attraction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trips</td>
<td>Firms</td>
<td>Trip Rate</td>
<td>Trips</td>
<td>Firms</td>
</tr>
<tr>
<td>Ontario</td>
<td>56,423</td>
<td>9,097</td>
<td>6.20</td>
<td>54,965</td>
<td>9,056</td>
</tr>
<tr>
<td>Rest of Canada</td>
<td>83,373</td>
<td>14,153</td>
<td>5.89</td>
<td>83,116</td>
<td>14,373</td>
</tr>
<tr>
<td>U.S.A.</td>
<td>58,507</td>
<td>25,161</td>
<td>2.33</td>
<td>59,716</td>
<td>25,479</td>
</tr>
</tbody>
</table>

For most industries, the trip rates for Ontario and the rest of Canada were similar. However, the Ontario trip rates exhibited issues caused by small sample sizes for the primary industries (agriculture, forestry, fishing, and mining) since they are under-represented (as shown previously in Figure 4-2). Therefore the Canadian trip rates (shown in Table 4-2) are adopted here since they provide a larger sample size for accurate representation.
Table 4-2: Initial trip rates for Canadian firms by industry

<table>
<thead>
<tr>
<th>Industry</th>
<th>Production</th>
<th></th>
<th>Attraction</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trips</td>
<td>Firms</td>
<td>Trip Rate</td>
<td>Trips</td>
</tr>
<tr>
<td>Agri., Forestry &amp; Fishing</td>
<td>1,498</td>
<td>285</td>
<td>5.26</td>
<td>1,466</td>
</tr>
<tr>
<td>Mining</td>
<td>778</td>
<td>133</td>
<td>5.85</td>
<td>816</td>
</tr>
<tr>
<td>Construction</td>
<td>9,697</td>
<td>1,925</td>
<td>5.04</td>
<td>9,827</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>23,900</td>
<td>3,821</td>
<td>6.25</td>
<td>23,556</td>
</tr>
<tr>
<td>Transportation</td>
<td>38,173</td>
<td>3,691</td>
<td>10.34</td>
<td>37,770</td>
</tr>
<tr>
<td>Wholesale Trade</td>
<td>12,416</td>
<td>1,931</td>
<td>6.43</td>
<td>12,461</td>
</tr>
<tr>
<td>Retail Trade</td>
<td>22,865</td>
<td>4,906</td>
<td>4.66</td>
<td>22,333</td>
</tr>
<tr>
<td>Services</td>
<td>22,291</td>
<td>4,956</td>
<td>4.50</td>
<td>22,116</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>139,796</strong></td>
<td><strong>23,250</strong></td>
<td><strong>6.01</strong></td>
<td><strong>138,081</strong></td>
</tr>
</tbody>
</table>

Values in this table pertain to one month of data

The total trip rate was also examined at the zonal (Ontario census division) level to determine if the trip rates are consistent over space. Figure 4-4 presents the relationship between the number of trips in the GPS sample and the subsequent trip production rate for each Ontario census division. It should be noted that five outliers (out of 49 zones) were removed from the plot, including three points with low total trips but very high trip rates (above 12) and two points with very high total trips but reasonable trip rates in line with the curve in Figure 4-4.

The trend line suggests that as the number of trips encountered for a given zone (i.e. the sample size) increases, the trip rate generally increases as well. However, the relationship itself is non-linear – as the number of trips increases, the trip rate increases at a slower pace. The trend suggests a general convergence of the trip rates approaching 7 trips per firm. The latter value also suggests that the sample derived trip rate may underestimate the actual trip rate of firms since Table 4-2 shows an average trip rate of
approximately 6. However, the under-estimation issue is handled through an optimization approach as will be described later in Section 4.4.6.

Figure 4-4: Total trip count and trip rate relationship for Ontario census divisions

4.4.2 Step 2 – Expanded Trip Totals

In Step 2 (with respect to Figure 4-3), the trip rates per firm by industry given in Table 4-2 are multiplied by the total firm counts (derived from an InfoCanada firm dataset) at the zonal level. This creates an expanded aggregate trip total that adjusts the industry proportions based on the frequency of firms in each zone. The result of this initial expansion is a set of production and attraction trip counts per zone by industry. The zones are based on Census Divisions in Ontario as defined earlier in Chapter 3.

4.4.3 Step 3 – Trip Distribution Using the IPF Method

Since the distribution of trips between origin and destination was previously found to be biased towards longer distance trips in the GPS dataset, this pattern was not utilized to disaggregate the production and attraction totals. Instead, the pattern of distribution from
the MTO 2006 CVS was used to create the origin-destination matrix. This distribution from
the CVS is based on roadside intercept surveys of trucks travelling across major highway
corridors in Ontario. The data is then expanded by MTO to account for biases, such as
double counting from multiple survey locations.

The iterative proportional fitting method (IPF)\(^1\) was applied to match the expanded
aggregate production and attraction totals (from Step 2) while preserving the underlying
spatial interaction pattern (OD seed matrix) derived from the CVS data. A summary of the
IPF inputs and outputs is provided in Figure 4-5.

\[^1\] IPF is also known in the transportation literature as the Fratar or Furness method (Ortuzar and
Willumsen 2011). The method has been widely used to calculate a new state of an Origin-
Destination (OD) matrix that conforms to known marginal rows and columns using an
existing (i.e. seed) OD matrix.

\[ T_{ij}^n = T_{ij}^{n-1} \times \frac{O_i^f}{O_i^{n-1}} \]  
(Eq. 4-2)

Figure 4-5: IPF method inputs and output

The IPF method takes the initial inputs and iterates until the initial seed matrix
matches the aggregate totals of production and attraction by zone. The first equation in the
IPF is provided below:
where \( T_{ij}^n \) are the resulting trips for each cell of the trip distribution matrix (with \( i \) rows and \( j \) columns) at the given iteration \( n \), and \( O_l^f \) are the final aggregate totals for the trip origins (productions) expected when the IPF procedure is complete. The right term for the equation, \( O_l^f / O_l^{n-1} \), calculates a proportional error of the trip origins based on the final aggregate total \( (O_l^f) \) and the aggregate totals from the previous iteration \( O_l^{n-1} \). This term is then applied to adjust the \( T_{ij} \) values from the previous iteration \( T_{ij}^{n-1} \). A similar calculation is then performed with the trip destinations (attractions) in the equation below:

\[
T_{ij}^{n+1} = T_{ij}^n \times D_j^f / D_j^n
\]

(Eq. 4-3)

where \( T_{ij}^{n+1} \) are the resulting trips for the next iteration \((n+1)\), \( T_{ij}^n \) are taken from the current iteration \((n)\) calculated in Equation 4-2, \( D_j^f \) are the aggregate totals for the trip destinations (attractions) expected when the IPF procedure is complete, and \( D_j^n \) are the current aggregate trip destination totals.

The IPF procedure is completed if the current values in the trip distribution matrix \((T_{ij})\) have converged compared to previous iterations. Convergence is achieved when the marginal row and column totals of the latest \( T_{ij}^{n+1} \) conforms to the target trip productions \( O_l^f \) and attractions \( D_j^f \) values (for all \( i \) and \( j \)). If convergence has not been achieved, Equations 4-2 and 4-3 are performed again with new iterations until a specified convergence criterion is met. More information on the IPF approach can be found in Lomax and Norman (2016).

4.4.4 Step 4 – Shortest Path Routes

To determine the suitability of the data obtained from the initial expansion, the resulting traffic flows of trucks are compared with point survey data along major highways
The estimated traffic flows emerging from the origin-destination results are based on an all-or-nothing traffic assignment between the 49 zones (Ontario census divisions) as shown in Figure 4-6 which are calculated using the Network Analyst extension of ArcGIS software. The free flow travel time was used for this purpose since the primary truck routes between these zones are typically large capacity highways that will also be utilized under congested conditions.

Figure 4-6: Shortest path routes between Ontario zones

4.4.5 Steps 5 / 6 – Allocation of Routes to Survey Locations

A relationship was developed between the shortest path routes for the 49 census divisions and the MTO survey points located across Ontario. This was done by determining the routes that pass along each survey point. Using this relationship, the origin-destination trips (from Step 3) were assigned to the appropriate routes. Next, the traffic volume was
further assigned to each survey station for comparison. A simple example of this process can be seen in Figure 4-7 with 3 zones and 3 origin-destination pairs. In this example, there are two routes that pass through station 2 (zone 1 to 3 and zone 2 to 3), therefore the estimated total trips that pass by the station include the trips from these routes (300 + 50). The results of this comparison for the Ontario network (in Figure 4-6) indicate that 77% of the total trips at the CVS survey stations are accounted for by the current expanded totals of the GPS sample trips.

**Figure 4-7: Correspondence between OD trips and survey stations**

<table>
<thead>
<tr>
<th>Zone i to Zone j</th>
<th>Trips ( T_{ij} )</th>
<th>Survey Stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 2</td>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>1 to 3</td>
<td>300</td>
<td>1, 2</td>
</tr>
<tr>
<td>2 to 3</td>
<td>50</td>
<td>2</td>
</tr>
</tbody>
</table>

4.4.6 Step 7 – Expansion Factor Optimization

The current totals from the GPS sample can be further expanded to match the trips observed by the CVS survey stations. In this case, the CVS data is based on 2006 vehicle counts performed at the stations shown in Figure 4-6. To accomplish this task, a non-linear optimization problem is formulated, where the objective function minimizes the total error between the CVS survey station totals and the traffic flows derived from the expanded GPS trip totals. The non-linearity requirement arises from the many-to-many relationship between OD zones and survey stations. As such, a single weighting multiplier value is
introduced to adjust all GPS trip totals simultaneously. The optimization is designed as follows:

Minimize: \( \varepsilon = \sum_{s}^{n} |T_{s,CVS} - wT_{s,GPS}| \)

Subject to: \( w \geq 0 \)

where \( \varepsilon \) is the total error to be minimized and \( w \) is the variable multiplier adjusted in the algorithm. \( T_{s,CVS} \) and \( T_{s,GPS} \) are the trip totals of survey station \( s \) from the CVS data and the GPS data, respectively, for all \( s=1, 2, \ldots, n \ (n = 45) \) survey stations located on at least one shortest path route. The optimization resulted in a final multiplier value of 1.27. This multiplier expands the origin-destination data derived from GPS trips a second time to reach a final total that corresponds to actual traffic totals as closely as possible.

A scatterplot showing the final CVS totals and expanded GPS totals is provided in Figure 4-8. The graph indicates a very strong one-to-one relationship between the two trip sets with a linear trend line slope of 1.01. Furthermore, the correlation between the two sets of data is 93.9%. A map of the errors suggested by Figure 4-8 are plotted in Figure 4-9. This map shows that the Toronto and Hamilton areas exhibit a higher actual total measured from the CVS data, while areas primarily north east of Toronto experience higher GPS totals. This can be expected due to the large number of intra-zonal (urban) trucks that are not accounted for by the GPS trips.
This chapter identified two types of bias, industry and distance, found in the GPS based sample of truck trips derived in Chapter 3. A method was established to remove the industry bias using trip rates and expanding by the population of firms in a given zone. In
addition, distance bias was accounted for by utilizing the IPF method to match total estimated zonal production and attraction (from the first expansion) while maintaining the origin-destination patterns obtained from the 2006 CVS survey created by MTO. A second expansion was then applied by optimizing the expanded GPS totals with the truck totals from survey station points located along major Ontario routes. The novel method introduced in the chapter provides a dual purpose of reducing bias in the dataset while simultaneously expanding the quantity of trips to match observed volumes.

While the 2006 CVS data was used for the second expansion, 2013 data has been prepared by MTO. Based on a simple comparison of the two datasets, it is expected that a truck trip increase of 40% between Ontario census divisions (102,175 trips in 2013 compared to 72,870 trips in 2006). In such a case, the 2006 multiplier value of 1.27 may be increased by 40% to 1.74. However, the final value may be slightly different if the 2013 trip counts did not increase at each location proportionally. The utilization of a single multiplier value for the second expansion has a clear advantage since the original expansion factors (trip rates by industry) can be combined with the optimized factor from the second expansion. For example, the manufacturing production trip rate of 6.25 per firm (from Table 4-1) and the second expansion factor of 1.27 would become 7.94 (6.25 × 1.27). The simplicity of a single factor for each industry type ensures that these trip generation rates are easily applicable in the Canadian context. The trip rates determined here work well on a macroscopic level with large zones, but would not be as appropriate for small zones where a single trip rate per firm by industry ignores variability in firm size and production outputs.

The total trip productions and attractions generated from the analysis provided a better representation of truck trips in Ontario compared to the original sample while closely
matching the aggregate totals observed on the road network. However, the microscopic
behaviour of individual trips is lost at an aggregate level. To retain the travel behaviour of
vehicles, the original sample can be used to synthesize a full population of trips by using
methods such as combinatorial optimization (Ryan et al., 2009). In such a case, the
synthesis algorithm can be used to ensure that the aggregate zonal totals by industry type
are maintained. Such a method has been applied before for expanded trip rates. For
example, Goulias et al. (2014) used population synthesis to expand a household survey in
California. After the trips are synthesized, this data can then be used in microscopic
transportation models (such as truck tours) without the biases inherent in the original GPS
sample.

While the expansion performed in this chapter utilized a single variable for the
expansion, this could easily be adapted into a multivariate expansion process. The trip rates
provided by industry in Table 4-2 could be turned into adjustable expansion variables
which would allow for separate expansions for each industry category. Moreover, certain
zones (or origin-destination pairs) could have a separate adjustment factor in the expansion
process. This would be useful in the event of heterogeneity across zones where some areas
exhibit a higher concentration of trips and correspondingly larger trip rates.
4.6 Chapter 4 References


Goulias, K., Ravulaparthy, S., Konduri, K., Pendyala, R. (2014) Using synthetic population generation to replace sample and expansion weights in household surveys for small area estimation of population parameters, *Compendium of the Transportation Research Board 93rd annual conference*.


5.1 Introduction

The movement of goods in the Canadian province of Ontario is an integral part of its economy. Ontario imports and exports each comprise roughly one third the provincial gross domestic product (GDP). A large proportion of this trade is funneled through several bridges at the international border between Ontario and the U.S. carrying 90% of Ontario’s international truck freight with the U.S. (Anderson, 2012). From the Canadian side, this includes the Queenston-Lewiston and Peace Bridges near Niagara Falls, Ambassador Bridge in Windsor, and Blue Water Bridge in Sarnia. The Ambassador Bridge has the largest number of trucks crossing between Canada and the U.S. with over 2.3 million crossings per year (PBOA, 2017). Many of the trucks crossing these bridges travel between freight hubs where goods may be consolidated.

In Canada, the Peel Region within the Greater Toronto Area stands as one of the largest freight hubs where over $1 billion dollars of goods are transferred every day (Peel Region, 2012). A considerable amount of those handled goods flow to/from the Chicago freight hub. The latter is one of the largest hubs in North America. The international trips observed for trucks travelling between these locations can be aggregated into two viable options as shown in Figure 5-1. The first option is a northern route that uses the Blue Water Bridge to connect Highway 402 in Sarnia with Interstate 94 in Port Huron. The
second option is a southern route that crosses the Ambassador Bridge between Highway 401 in Windsor to connect with several major interstate highways in Detroit.

![Map showing northern and southern routes between Toronto and Chicago](image)

**Figure 5-1: Northern and southern routes between Toronto and Chicago**

The travel times shown in Figure 5-1 represent the five percentile trip time derived from GPS data that were utilized in the analysis of this thesis for the month of March 2013. As can be seen from the five percentile crossing times in Figure 5-1 (the minimum travel times are not presented to avoid extreme outliers), the two border crossings alternatives present similar opportunities when unimpeded. In addition, the average crossing times are likewise similar for the two routes. This presents a unique case since the route choice of a given user (i.e. carrier) is normally based on travel time savings. In the absence of a clear time differential between these two choices, this chapter seeks to answer two questions:

1. what are the factors that give rise to the choice of border crossing location faced by each truck travelling along this corridor

2. how much does the level of service of a border crossing influence its attractiveness?
A binary discrete choice model is specified and estimated to explain the observed choice behavior. While the wait delay at a border crossing can certainly have a detrimental effect on the economy (Park et al., 2014), it is expected that this may also influence the choice decision for a particular route. Therefore the hourly crossing time at the border was one of the primary factors included in the choice model with an expectation that a higher delay at a given crossing will reduce the propensity of choosing the route facilitated by that crossing. To the author’s knowledge, the analysis presented in this chapter is novel and has not been conducted in previous research.

5.2 Background on Route Choice Modelling

Studies pertaining to the route choice characteristics of freight are scarce compared to passenger route choice behavior (Feng et al., 2013). Several reasons can be attributed to the lack of studies on the topic including: (1) confidentiality/liability concerns that make freight data more difficult to obtain, and (2) the supply-chain process can be very complex when moving goods between manufacturing origins and final customer destinations. In addition, each stage of this process may have different or multiple decision makers. For example, the retailer, distributor, or freight forwarder who may organize the shipments (Feo-Valero et al., 2011). Also, the shipper may have different decision makers involved in route choice planning including company planners/dispatchers and the drivers themselves (Feng et al., 2013). Passenger travel is comparatively simpler by contrast given that the occupants of the vehicle are typically the sole decision makers.

Since freight data for route choice modelling is difficult to obtain, most of the existing studies rely on stated preference surveys (Nielsen, 2004; Arentze et al., 2012). In
other cases, the modeller may obtain data from road-side intercept surveys where the respondent completes a questionnaire on site about their route choice preferences (Quattrone and Vitetta, 2011). Finally, passive data can be used from sources such as global positioning system (GPS) data-loggers that record pings identifying the location and time of a traveling vehicle (Nielsen, 2004; Quattrone and Vitetta, 2011). The last option, which is used in this thesis, is becoming increasingly popular as the modern world becomes technologically integrated and dependent on geographic information. This technology has gained noticeable interest in the transportation field in recent years due to the large amount of travel pattern data it produces. The generated GPS records can be classified as Big Data based on the volume, velocity and variety of the provided information (Chen and Zhang, 2014). For instance, the GPS truck database in this thesis for March 2013 is approximately 18.56 gigabytes in terms of its volume when stored in basic ASCII format. Furthermore, our database is based on a variety of transponders/trucks that were involved in generating the GPS data.

The large amount of complexity for freight shipments has led to a heightened attentiveness among modellers towards heterogeneity. This includes heterogeneity across commodity type (Feo-Valero et al., 2011), vehicle size (Feng et al., 2013), intermodal availability (Patterson et al., 2007), and the value of time (Feo-Valero et al., 2011; Nielsen, 2004). For this reason, emergent discrete choice models that are better suited towards capturing heterogeneity when dealing with route choice behavior include the mixed logit and latent class models (Feng et al., 2013). In addition, there are two other prominent issues in route choice modeling including (1) the generation of perceived route choices given that users do not have full information about all alternatives, and (2)
correlations and overlap among potential routes between a given origin-destination pair (Frejinger and Bierlaire, 2007; Prato, 2009). The issue of route substitutions/overlap has led to the use of specialized models such as the C-logit (Cascetta et al., 2002) and path size logit (Ben-Akiva and Bierlaire, 1999). However, those problems are more applicable to urban travel in which the number of alternative routes is relatively large.

The model presented in this chapter does not suffer from those two problems since the alternatives for border crossing along the Toronto-Chicago corridor are fixed and do not overlap. Therefore, a binomial logit model can be employed to study the route choice with the two bridges as potential alternatives. To account for unobserved heterogeneity among the modeled trucks, the mixed logit model can also be employed in this discrete choice problem.

5.3 Data

The primary data used in the analysis is comprised of global positioning system (GPS) ping records that were generated from the movement of trucks along the Toronto-Chicago corridor during the month of March 2013. Those data were extracted from a larger dataset provided by Transport Canada. Each GPS ping provides a data record with a corresponding truck and carrier ID along with a time stamp and the geographic coordinates of the truck as shown previously in Table 1-1. While these pings are often recorded every 5 to 15 minutes for a given truck, the time lapse can occasionally increase up to several hours or higher. This can lead to issues involving sparse data if the pings are not frequent enough to determine the traversed routes.

The full dataset for the month of March 2013 includes 101.6 million individual GPS pings belonging to 40,650 trucks. These points were then processed to only include
direct trips between Toronto and Chicago. A direct trip is defined here as a set of GPS pings for a given truck that exhibits a stop of 15 minutes or greater in both Toronto and Chicago with no unusual movements away from the destination. Typically, unusual movements would suggest an intermediate stop by the truck, which could potentially influence the route choice. Such behavior would negatively impact the statistical integrity of the route choice model. The 15 minute time window was partially selected to reduce the probability of identifying a stop caused by traffic congestion. Although small deliveries may take less than 15 minutes to complete, their occurrence is less frequent for trucks travelling between major hubs such as Toronto and Chicago.

The number of direct movements along the studied corridor resulted in 3,111 trips (251,643 GPS pings). There were 1,264 trips (103,787 GPS pings) for trucks heading from Toronto to Chicago and 1,847 trips (147,856 GPS pings) for trucks heading in the opposite direction. The total number of trips was later reduced to 1,389 due to sparse data. The respective shares of trips crossing the Ambassador Bridge and Blue Water Bridge in our data sample are 60.3% (837 out of 1,389) and 39.7% (552 out of 1,389). Data from the Public Border Operators Association (2017) during the March, 2013 time period suggests similar proportions for all trucks crossing the Ambassador and Blue Water bridges at 61.2% (195,836 out of 320,137) and 38.8% (124,301 out of 320,137), respectively. Therefore the sample of GPS derived trips between Toronto and Chicago is similar to the total frequency of trucks utilizing these crossings.

To control for the effect of industry type on the crossing choice behavior, the modelled trips were linked to the nearest business establishment at the destination. Although the nearest business is most likely the true destination of the trip, some
adjustments were applied to destination locations in close proximity to service stations. In essence, if a destination has a service station within a radius of 200 meters, then the trip is not linked to any establishment and no industry is assigned. This treatment was necessary to minimize the chance of generating false positives in which a service stop is treated as a final destination (i.e. delivery) stop. Next, trips linked to establishments are assigned the Standard Industrial Classification (SIC) code found on record. Two firm datasets were acquired and utilized for this task including (1) DatabaseUSA to provide data on Chicago firms and (2) InfoCanada to provide data on Toronto firms. More information on this processing can be found in Chapter 2.

5.4 Methods of Analysis

5.4.1 Mixed Logit (MXL) Model

A mixed logit (MXL) model was utilized to explain the selection of border crossing for trucks moving directly between Toronto and Chicago. Modelled across two possible alternatives (northern and southern routes in Figure 5-1), the MXL is mainly employed to account for unobserved heterogeneity in the choice behavior. While the driver of a vehicle is typically the decision maker in the case of personal travel, this may not be true when dealing with freight truck movements. The organizational structure of a freight carrier may dictate who performs the route choice decision making. This may be undertaken by a number of heterogeneous agents including the driver of the vehicle, the dispatcher for a given carrier firm, or owner of the shipped goods. However, the utilized GPS data did not include information regarding the decision makers themselves. As such, the MXL has an advantage over the conventional binomial logit model because it is
capable of capturing some of the unobserved heterogeneity among the various decision makers (i.e. taste variation).

In the MXL, a decision maker $t$ will choose route/crossing $r$ that provides the greatest utility from a feasible set of discrete alternatives $R$. The choice probability for decision maker $t$ selecting $r$ can be formulated as follows (Train, 2009):

$$P_{tr} = \int P_{tr/\beta_t} \phi(\beta_t|\mu, \sigma)d\beta$$  \hspace{1cm} (Eq. 5-1)

where $P_{tr/\beta_t}$ is the logit probability of decision maker $t$ selecting route $r$ conditional on a given $\beta_t$ value. The probability of having a particular $\beta_t$ value can be obtained by drawing from a known probability density function $\phi(.)$ that has mean $\mu$ and standard deviation $\sigma$. According to Train (2009), the choice probability $P_{tr}$ in equation 5-1 can be thought of as the weighted probability of $P_{tr/\beta_t}$ across all possible $\beta_t$ values. $\phi(.)$ in equation 5-1 represent the weights associated with $\beta_t$. Typically, $\phi(.)$ is assumed to follow the normal distribution although other functional forms such as the lognormal have been used. Parameters $\mu$ and $\sigma$ of $\phi(.)$ are estimated based on the distribution assumed for the latter. The mixed choice probability $P_{tr}$ conforms to the ordinary logit (ORL) model if the estimated $\sigma$ is not significant (i.e. cannot be differentiated from 0).

The utility of the logit model is a unit-less representation of the level of satisfaction provided by routing through one of the two border crossings. In the case of private firms, this satisfaction will generally arise from the route that maximizes profits by minimizing travel time. However, travel time between Toronto and Chicago is roughly the same along the two modeled routes. Therefore, it is expected that characteristics pertaining to the border crossings themselves and/or the decision makers are responsible for the revealed crossing choices.
5.4.2 Explanatory Variables

The independent variables used in the specification of the observed utilities are based on \textit{a priori} expectations. A summary of these variables are included in Table 5-1 below.

\begin{table}[h]
\centering
\small
\begin{tabular}{ll}
\hline
\textbf{Variable} & \textbf{Description} \\
\hline
$CTIME_r$ & Average border crossing time for a given hour of the day for crossing $r$ (by direction) \\
$TI_n$ & A dummy variable for a truck crossing during a specific time interval $n$ (1 if true; 0 if false) \\
$D_j$ & A dummy variable for a truck heading in a certain direction $j$ ($j = \text{Toronto or Chicago}$); (1 if true; 0 if false) \\
$CARRIERS_r$ & A dummy variable for carriers with a very high preference for a particular crossing $r$ (1 if true; 0 if false) \\
$PASTCHOICE_r$ & A variable containing the sum of previous crossings for a given truck at crossing $r$ \\
$INDUSTRYSIC$ & A dummy variable for trips associated with a particular SIC industry type (1 if true; 0 if false) \\
$DAY_d$ & A dummy variable for trips crossing the border on a given day $d$ of the week (1 if true; 0 if false) \\
\hline
\end{tabular}
\caption{Variables used in the crossing choice model}
\end{table}

The primary expectation for the average crossing time $CTIME_r$ variable is that the decision maker is aware of the time of day they will reach the border and will adjust their decision according to hourly variations in crossing statistics. The average crossing time for a given hour of the day over the 31 days of March 2013 was calculated using observed crossing times from the full set of GPS data. This included over 17,000 observations for the Ambassador Bridge and 9,000 observations for the Blue Water Bridge. Hourly variations for the average crossing time are shown in Figure 5-2 for the two crossings. It is expected that trucks will prefer a lower average crossing time. For example, a truck crossing toward the U.S. between 6:00 PM – 7:00 PM would favor the 15.6 minute average crossing time at the Blue Water Bridge compared to 22.4 minutes at the Ambassador Bridge.
Figure 5-2: Average times at the Ambassador (top) and Blue Water (bottom)

The proportions of hourly crossing volumes for trucks are shown in Figure 5-3. These proportions are derived from the fraction of crossing occurrences for a given hour to the total number of crossings observed from the GPS dataset during the month of March 2013. The proportions varied by the time of day and were higher or lower for a given crossing during certain time intervals. For instance, the proportions of hourly trips going to Canada via the Blue Water Bridge were higher than their Ambassador Bridge counterpart for the 5:00 AM – 1:00 PM time interval. Similarly, the proportions of hourly trips crossing the Ambassador Bridge to Canada were higher for the 2:00 PM to 10:00 PM time interval with the exception for 6:00 PM. The same phenomenon was detected in the case of the trips crossing to the United States. More specifically, the hourly proportions of trips crossing to the US via the Blue Water Bridge were higher during the 7:00 AM to 3:00 PM period. On the other hand, the hourly proportions of trips crossing
the Ambassador Bridge to the US were higher during the 7:00 PM to 4:00 AM period. To control for these effects, time interval variables $T_{In}$ for four periods $n$ ($n = 5:00$ AM – 1:00 PM; 2:00 PM – 10:00 PM; 7:00 AM – 3:00 PM; 7:00 PM – 4:00 AM) were introduced and interacted with the direction variables $D_j$ in the model.

![Exploration of the data revealed that trucks belonging to certain carriers (as identified by their carrier id) tend to choose the same crossing on a regular basis during the month of March 2013. As such, a CARRIERS$_r$ variable was included in the model to control for this effect. Such behavior is suggested to be the outcome of trucks being influenced by dispatchers who have a predisposition towards one crossing over the other. Another set of variables, PASTCHOICE$_r$, were added to control for the effect of correlations arising from multiple observations for a single vehicle. For a given

Figure 5-3: Crossing volumes towards Canada (top) and the U.S. (bottom)
observation, \( PASTCHOICE_A \) and \( PASTCHOICE_B \) measure the number of times the given truck had previously used the Ambassador Bridge and Blue Water Bridge, respectively. A separate model with this measure of serial correlation on the choice of crossing was included to establish the large influence it has on the final model fit \( (\rho^2) \). In addition, industry specific variables \( INDUSTRYSIC \) were introduced in the model to control for the effect of industry type on crossing choice.

Figure 5-4 presents the breakdown of daily crossings between the two bridges. Overall, the market shares of the observed 1,389 trips between Toronto and Chicago during the month of March 2013 were 60% and 40% for the Ambassador Bridge and Blue Water Bridge, respectively. A glance at Figure 5-4 indicates that the daily market shares were consistent and in line with the overall 60/40 distribution. However, such a split was not observed for the Monday and Thursday trips. An excess of 13% in favor of the Ambassador Bridge was observed for Monday while an excess of 10% in favor of the Blue Water Bridge was observed for Thursday. A closer look at the excess market shares for the Ambassador Bridge every Monday of March 2013 suggests a consistent pattern (Week 1 = 13%, Week 2 = 17%, Week 3 = 11% and Week 4 = 13%) that is in line with the aggregated pattern. By comparison, an examination of the excess market shares for the Blue Water Bridge every Thursday of March 2013 did not show a similar pattern (Week 1 = 15%, Week 2 = 8%, Week 3 = 16% and Week 4 = 8%). The \( DAY_d \) variables were introduced in the model to control for the market share excesses observed for the Monday and Thursday trips.
Finally, an extra variable in the initial model estimation contained the average crossing time two hours in advance of the actual crossing time and date. This was included to account for the possibility of switching bridges mid-route. This variable was not found to be significant in the model and was later dropped from the final results.

5.5 Results

The variables presented in the previous section (Table 5-1) were employed in the specification and estimation of binary discrete choice models using NLOGIT 4.0 statistical software. Table 5-2 presents the results for three models: an ordinary logit (ORL); mixed logit (MXL); and mixed logit with past truck choices. The latter two models were estimated using 500 random Halton draws. The choice models in all cases...
are well behaved in terms of the expected signs. As shown in Table 3, both the ORL and MXL models have acceptable adjusted-$\rho^2$ values of 0.127 and 0.128, respectively. However, including serial correlation of past choices for a given truck substantially increased the adjusted-$\rho^2$ value to 0.182. The corresponding parameters for $PASTCHOICE_A$ and $PASTCHOICE_B$ were highly significant. These results suggest that the previous choices of a vehicle are a very large indicator of future decisions.

**Table 5-2: Discrete choice model results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Choice</th>
<th>Binomial Logit</th>
<th></th>
<th>Mixed Logit</th>
<th></th>
<th>Mixed Logit with Past Truck Choices</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Parameter</td>
<td>T-stat</td>
<td>Parameter</td>
<td>T-stat</td>
<td>Parameter</td>
<td>T-stat</td>
</tr>
<tr>
<td>$CONSTANT$</td>
<td>A</td>
<td>-0.483</td>
<td>-1.038</td>
<td>-0.442</td>
<td>-0.932</td>
<td>-0.474</td>
<td>-0.96</td>
</tr>
<tr>
<td>$CTIME_A$</td>
<td>A</td>
<td>-0.055</td>
<td>-2.366**</td>
<td>-0.071</td>
<td>-2.746**</td>
<td>-0.069</td>
<td>-2.532**</td>
</tr>
<tr>
<td>$CTIME_B$</td>
<td>B</td>
<td>-0.120</td>
<td>-4.090**</td>
<td>-0.138</td>
<td>-4.223**</td>
<td>-0.148</td>
<td>-4.34**</td>
</tr>
<tr>
<td>$TI_{14,22} \times D_{Toronto} (\mu)$</td>
<td>A</td>
<td>0.314</td>
<td>1.843*</td>
<td>0.621</td>
<td>1.833*</td>
<td>0.189</td>
<td>1.048</td>
</tr>
<tr>
<td>$TI_{14,22} \times D_{Toronto} (\sigma)$</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>1.786</td>
<td>1.999**</td>
<td>0.033</td>
<td>0.021</td>
</tr>
<tr>
<td>$TI_{5,13} \times D_{Toronto} (\mu)$</td>
<td>B</td>
<td>0.428</td>
<td>2.010**</td>
<td>0.538</td>
<td>1.647*</td>
<td>0.451</td>
<td>1.845*</td>
</tr>
<tr>
<td>$TI_{5,13} \times D_{Toronto} (\sigma)$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>2.192</td>
<td>1.907*</td>
<td>0.877</td>
<td>0.829</td>
</tr>
<tr>
<td>$CARRIERS_A$</td>
<td>A</td>
<td>2.619</td>
<td>5.618**</td>
<td>2.765</td>
<td>5.553**</td>
<td>2.391</td>
<td>5.084**</td>
</tr>
<tr>
<td>$CARRIERS_B$</td>
<td>B</td>
<td>3.514</td>
<td>5.845**</td>
<td>4.188</td>
<td>5.042**</td>
<td>3.383</td>
<td>5.411**</td>
</tr>
<tr>
<td>$PASTCHOICE_A$</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.282</td>
<td>4.466**</td>
</tr>
<tr>
<td>$PASTCHOICE_B$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.729</td>
<td>7.792**</td>
</tr>
<tr>
<td>$DAY_{Mon}$</td>
<td>A</td>
<td>0.421</td>
<td>1.889*</td>
<td>0.446</td>
<td>1.888*</td>
<td>0.375</td>
<td>1.592</td>
</tr>
<tr>
<td>$DAY_{Thu}$</td>
<td>B</td>
<td>0.409</td>
<td>2.456**</td>
<td>0.493</td>
<td>2.610**</td>
<td>0.410</td>
<td>2.297**</td>
</tr>
<tr>
<td>$INDUSTRY_{35}$</td>
<td>A</td>
<td>2.262</td>
<td>3.055**</td>
<td>2.703</td>
<td>2.898**</td>
<td>2.296</td>
<td>3.018**</td>
</tr>
<tr>
<td>$INDUSTRY_{47}$</td>
<td>A</td>
<td>0.587</td>
<td>1.694*</td>
<td>0.696</td>
<td>1.627</td>
<td>0.524</td>
<td>1.404</td>
</tr>
<tr>
<td>$INDUSTRY_{56}$</td>
<td>B</td>
<td>3.058</td>
<td>2.957**</td>
<td>3.347</td>
<td>2.883**</td>
<td>3.363</td>
<td>3.186**</td>
</tr>
<tr>
<td>$INDUSTRY_{79}$</td>
<td>B</td>
<td>0.856</td>
<td>1.585</td>
<td>0.918</td>
<td>1.607</td>
<td>1.051</td>
<td>1.941*</td>
</tr>
</tbody>
</table>

Log-Likelihood (only constants) -933.3 -933.3 -933.3
Log-Likelihood (final) -806.9 -804.9 -754.1
$\rho^2$ 0.136 0.138 0.192
Adjusted $\rho^2$ 0.127 0.128 0.182

$A = $ Ambassador Bridge (837 records); $B = $ Blue Water Bridge (552 records)

** statistically significant to 95%; * statistically significant to 90%

According to the results, the parameters of the crossing time variables $CTIME_r$ ($r = A$ for Ambassador Bridge and $B$ for Blue Water Bridge) are statistically significant and retain the correct negative signs. This suggests that trucks will tend to avoid a border
crossing with higher travel times, other things being equal. The direct elasticity was calculated for both bridges and found to be higher for the Blue Water Bridge compared to the Ambassador Bridge with values of -1.231 and -0.476, respectively. For the Blue Water Bridge, this is a fairly elastic result suggesting that a 1% increase in the border crossing time at that location will reduce the probability of selecting this bridge by 1.2%.

A sensitive analysis was performed in NLOGIT using the ORL model to determine a 50-50 break-even point in the proportion of vehicles selecting each bridge. The original data has a 60% proportion favoring the Ambassador Bridge. Based on the performed simulations, a 50-50 split in vehicles is achieved if the average crossing time at the Ambassador Bridge is increased by 47% while all other variables are held constant. The results of this scenario are given in Table 5-3. To generate an opposite split in which 60% of the trucks would favor the Blue Water Bridge, the crossing time on the Ambassador Bridge has to increase by 95%. This suggests that a significant increase in the average border crossing time in the Ambassador Bridge would be required to draw away the majority of the trucks towards the Blue Water Bridge. In a hypothetical case the average crossing travel time on the Ambassador Bridge will have to increase by 750% to imitate the closure of that Bridge where all trucks go through the Blue Water Bridge. Conversely, the average crossing time of the Blue Water Bridge will have to increase by 400% to direct all traffic through the Ambassador Bridge. This is indicative of the role of the Ambassador Bridge in facilitating the movement of trucks along the Toronto-Chicago corridor.
Table 5-3: Sensitivity of Ambassador crossing times

<table>
<thead>
<tr>
<th>Border Crossing</th>
<th>Ambassador Bridge Avg. Crossing Time Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0%</td>
</tr>
<tr>
<td>Ambassador</td>
<td>60.3</td>
</tr>
<tr>
<td>Blue Water</td>
<td>39.7</td>
</tr>
<tr>
<td>Border Crossing</td>
<td>30%</td>
</tr>
<tr>
<td>Ambassador</td>
<td>53.8</td>
</tr>
<tr>
<td>Blue Water</td>
<td>46.2</td>
</tr>
</tbody>
</table>

The expectation that certain carriers will have a strong inclination towards a specific crossing appears to hold true. This is apparent through the positive and statistically significant parameters of the two variables $CARRIERS_A$ and $CARRIERS_B$ controlling for carriers holding a strong preference for the Ambassador Bridge and Blue Water Bridge, respectively. The $CARRIERS_A$ variable captured the preference of 5 individual carriers while $CARRIERS_B$ captured the preference of 4 specific carriers. Trucks pertaining to these carrier groups almost exclusively selected one crossing over the other while the remaining 112 carriers varied their route choice between the two crossing alternatives. The distinct preference exhibited by certain carriers could be caused by a familiarity with one of the bridge locations over the other. A different possibility is that some of these carrier companies might use custom brokers located near a particular bridge. A third possibility could be that trucks pertaining to these carriers specialize in servicing an exclusive type of industry (e.g. automotive) and as such would favor one crossing over the other. However, an examination of the industries serviced by the trucks suggest a very similar industry breakdown among the two groups of carriers. Similar distributions were also found for the full count of trucks using either crossing point between Toronto and Chicago.
Out of all the tested industry variables, four significant ones were kept in the model to capture the effect of heterogeneity for different types of goods. Trucks associated with two industries had a higher likelihood of selecting the Blue Water Bridge including retail apparel and accessory stores (\textit{INDUSTRY}_{56}) and construction contractors (\textit{INDUSTRY}_{17}), although the former was noticeably more significant. On the other hand, two other industries showed higher preference for the Ambassador Bridge including industrial and commercial machinery/equipment (\textit{INDUSTRY}_{35}) and transportation services (\textit{INDUSTRY}_{47}). It is worth noting that the transportation services industry does not relate to any given type of shipped goods. Instead, trips associated with those industries at the destination indicate that the truck stopped at some type of carrier yard or shipping depot. This would result in the goods typically being prepared to continue towards some other destination.

The $TI_n$ ($n = 7:00$ AM – 3:00 PM; 7:00 PM – 4:00 AM) variables for trips destined to Chicago were dropped from the final specification of the model due to a lack of statistical significance. However, other things being equal, trips destined to Toronto between 2:00 PM and 10:00 PM showed a strong preference towards the Ambassador Bridge. Also, trips occurring in the same direction between 5:00 AM and 1:00 PM favored the Blue Water Bridge. The estimates of the MXL model indicate that the parameters of the latter two time interval variables are random and can be drawn from a normal distribution. More specifically, the parameter associated with the variable $TI_{14-22 \times D_{Toronto}}$ can be drawn from a normal distribution $\mathcal{N}(\mu = 0.621, \sigma = 1.786)$, while the parameter of $TI_{5-13 \times D_{Toronto}}$ can be drawn from $\mathcal{N}(\mu = 0.538, \sigma = 2.192)$. The results of the randomized parameters suggest that trucks crossing the Ambassador Bridge between
2:00 PM and 10:00 PM do not possess the same preference (i.e. taste) for that crossing and time interval, other things being equal. The same could be said about the trucks crossing the Blue Water Bridge between 5:00 AM and 1:00 PM. The choice of border crossing was also influenced by day of the week as we expected based on Figure 5-4. Parameters for both variables $DAY_{Mon}$ and $DAY_{Thu}$ are significant and retain the correct positive sign. However, the $DAY_{Mon}$ variable is no longer statistically significant with 90% confidence in the third model (mixed logit with past truck choices). These findings suggest that the Ambassador Bridge was selected more frequently on Mondays while the Blue Water Bridge was proportionally more popular on Thursdays.

5.6 Conclusions

This chapter utilized GPS data corresponding to truck movements in North America to derive direct trips between Chicago and Toronto. The initial analysis found that there were two possible alternatives for crossing between Canada and the U.S. for the Toronto-Chicago route - a northern route through the Blue Water Bridge and a southern route through the Ambassador Bridge. A binary logit model was therefore introduced to capture possible factors influencing the selection of the two border crossings. The international crossings included in the study area are among the busiest border crossings in the world. As such, they play an important role in sustaining trade and ensuring economic prosperity for both Canada and the US. The results provided in this paper are novel as they make a direct contribution to the scarce literature on cross-border transportation.

The statistical analysis supported the hypothesis that a larger average crossing time during a given time of day at one border crossing will increase the proportion of
trucks selecting the other crossing. This confirms the expectation that trucks will select a route that avoids inefficiencies in crossing times. The model suggests that a 47% increase in the crossing time at the Ambassador Bridge would decrease its share of trucks to an even split with the Blue Water Bridge. From a practical perspective, the popularity of a second bridge crossing that is planned to be built between Detroit and Windsor (the Gordie Howe International Bridge) may have some dependency on the crossing time efficiency at the border plaza. This could occur despite new feeder highways that will slightly reduce the overall time to reach the border. Future modelling efforts could be undertaken in the future to study the selection of border crossing when the new bridge alternative becomes available. In such a case, a nested logit model could be introduced to group the two crossings between Windsor and Detroit together in the upper tier due to their proximity to each other. The results also revealed that some carriers have a very strong preference for one crossing over the other. Some carrier firms preferred the Ambassador Bridge while several others preferred the Blue Water Bridge. This could be due to the preferences of a dispatcher or other decision maker for a given firm controlling the routes of all their trucks.

The findings presented in this chapter offer a unique perspective on trucks crossing the international border between Canada and the U.S. A limitation to this study is that the discrete choice model is based on an assumption that the alternatives (i.e. border crossings) are independent of each other. However, the crossing times for both bridges are partially governed by the same organization suggesting some dependency between the two crossings. On the Canadian side, both bridges are controlled by the
Canadian Border Services Agency (CBSA) while the U.S. border inspections are conducted by the U.S. Customs and Border Protection (CBP).

While this study used data for a one month time period to estimate a cross-sectional model, more data is available. Future efforts could include estimates for other time periods in addition to time-series studies to measure the impact of seasonality on border crossing choices. Even though the total travel time is equivalent by route, the actual travel cost includes the time required to make the journey and the toll cost when crossing the border. While the toll information is easy to obtain, they typically vary by weight or axle. Therefore, extra information on truck size and weight would be a useful addition to the model in the future to account for toll fees. From a practical perspective, this type of model could then be integrated within a toll competitiveness model such as Li et al. (2014) which looked at optimizing toll prices between the Ambassador Bridge and a new anticipated border crossing between Windsor and Detroit. The results from this chapter could add to this competitive price model by expanding the model to include the Blue Water Bridge and adding truck crossing choice behavior. Including other crossings and scenarios is also possible but would require the addition of the total trip time as an explanatory variable to account for differences that commonly arise between routes.

It would also be beneficial to study the effect of route distances on the sensitivity of border crossing choice behavior. For example, a truck traveling on a longer trip may not be as affected by crossing delays since it will represent a smaller fraction of the total trip time. This effect of distance could be captured by identifying trips that span beyond the Toronto-Chicago origin-destination trip ends. The relationship between trip distance travel delay is a topic of interest in the next chapter (Chapter 6).
5.7 Chapter 5 References


6.1 Introduction

Transportation delays are a commonplace occurrence for passenger vehicles and trucks alike. In the latter case, delays can have a direct impact on a region’s economic performance due to the added costs incurred by producers. This burden is amplified when delays are unpredictable, resulting in an uncertain arrival time for truck deliveries. The delays associated with these deliveries can be attributed to a number of factors including congestion on major highways, custom clearance at border facilities, or unexpected events such as road accidents or inclement weather.

6.1.1 Main Objectives

To understand the nature of truck delays in North America, it is imperative to quantify the extent that it occurs on congested highway traffic and at the border (if a truck is exporting goods internationally. As such, this chapter adds to existing literature by presenting a new measure of reliability on roads and calculating this measure using big data extracted for one month in 2013 that represent the movements of over 30,000 Canadian owned trucks using GPS pings. These pings track the geographic position of trucks at different points in time. The primary objectives of this paper are twofold:

1. Examine Canada-U.S. border delays as a function of the total trip time. Other studies have measured border wait times before (Paselk and Mannering, 1994;
Goodchild et al., 2008; McCord et al., 2010; Lin et al., 2014), but the relative effect on the total trip has not been analyzed in detail to date.

2. Present an intuitive measure of delays occurring on road infrastructure as a proportion of the total trip time and apply the delay measure on a large dataset of Canadian trucks. The delays are further deconstructed into expected and unexpected delays due to the extra costs associated with the latter category.

6.1.2 Data Types

The opportunity to comprehensively study delays experienced on the road network is becoming increasingly available as technological innovations and their widespread adoptions occur. Historically, information on road users was obtained by phone/mail using surveys based on memory recalls or using intercept surveys where short roadside questionnaires are answered regarding the current trip. These traditional types of data are purposely-sensed, indicating that they are designed and implemented specifically for the purpose of gathering information on transportation patterns. The International Transport Forum (2015) categorizes transport data types as purposely-sensed, opportunistically-sensed, and crowd sourced.

By contrast to purposely-sensed data, opportunistically-sensed data is produced for some other purpose but later adapted to provide information to transportation researchers. GPS data for trucks (discussed in more detail later), is a prominent example of opportunistic data. Originally used by carrier fleets to coordinate the positions of their trucks, the GPS data can be utilized after some data cleaning/manipulation to provide vast amounts of information on trucking patterns to researchers. For example, numerous studies in the U.S., such as Zanjani et al. (2015) and Kuppam et al. (2014), have utilized
GPS data purchased from the American Transportation Research Institute (ATRI). The pervasiveness of cell phone usage can also be leveraged to provide opportunistic data as seen by studies such as Jiang et al. (2016) that utilized call detail records (CDR) from 1.92 million cell phone users over 6 weeks in the Boston area. Finally, the third type of information, crowd sourced data, makes use of content-sharing platforms such as transportation oriented applications (e.g., the Google Waze app), or apps typically unrelated to transportation (e.g., Facebook or twitter), where the locations/behaviours of individuals can be identified. Both of the latter two data types, opportunistic-sensed and crowd sourced, can provide immense amounts of information at a much lower cost since there are no setup or infrastructure costs.

6.1.3 Measures of Reliability

Along with the increasing availability of data to transportation analysts, detailed measures of reliability are being developed to utilize such data for the analysis of road segments. These measures have become important for evaluations of the performance of road links and the prioritization of transportation upgrades. In Canada, a survey conducted by the Transportation Association of Canada of provincial government performance measures was categorized based on six survey outcomes: safety, infrastructure preservation, sustainability, cost effectiveness, reliability, and mobility (TAC, 2006). In the U.S., performance measures applied to roads are a vital aspect of the MAP-21 funding initiative (FHWA, 2012), leading to a recent increasing interest in the subject.

Among the outcomes listed by the TAC (2006) survey, reliability is the most relevant to this paper. The survey describes several reliability measures including a level
of service, the percentage of vehicle kilometers spent in congested traffic and the total duration of highway closures. The related outcome of mobility is quantified using simple measures such as speed or volumes. However, the definition is sometimes be interchanged with reliability. For example, Wang et al. (2016) categorized reliability measures as 1) spot speed indicators for single points in time, and 2) travel time based indicators measuring the time reliability of a particular road segment. The former spot speed indicators are often labeled a mobility indicator as a determination of the instantaneous speeds of vehicles at a given location. The spot speed can be easily estimated based on distance and elapsed time between two consecutive measurements. Liao (2014) compared the spot speed with a space mean speed measure where the calculation is based on a specified distance of roadway. The study found that spot speeds exhibit standard deviation errors that are twice as large as the space mean speed, suggesting that the former measure may have issues with inconsistent results.

Several travel time reliability measures include direct percentile measurements (such as 90 percentile) and a buffer index representing the factor of time needed to meet a given percentile time (Wang et al., 2016). These measures have notably been introduced in the past for border crossing reliability comparisons (Anderson and Coates, 2010). Time based measures suggested by the American Association of State Highway and Transportation Officials (AASHTO) to meet the requirements of MAP-21 include a Truck Reliability Index (that is very similar to the buffer index) and Annual Hours of Truck Delay (AASHTO, 2012). Both of the aforementioned measures are based on a defined threshold level of acceptable travel speed as defined by a given agency. Such measures have been implemented in past studies (such as Liao, 2014), however, these
measures provide limited appeal to decision makers who may not fully understand the meaning behind them. One approach to provide easy interpretation is to apply a qualitative categorization to the travel characteristics. For example, Zhao et al. (2013) identified some road segments in Washington State as ‘Unreliable’, ‘Reliably Fast’, or ‘Reliably Slow’ based on the properties of a bimodal distribution of speeds generated for each road segment and time period. However, such categorization is not required if the performance measure itself is easy to interpret.

This chapter presents a more intuitive metric of observing delays as a proportion of the recorded total travel time of individual trips. This type of approach requires some type of vehicle probe data that observes an entire trip. However, the impact of delays as a function of the full trip duration is easier to interpret compared to delays at any single point in time or an index measuring variability. For example, a delay of 10 minutes at a single location is very substantial if the total trip only takes 20 minutes (50% delay) but may be less substantial if the total trip takes 2 hours (8.3% delay).

Another distinction between existing reliability measures and the proposed proportional trip delay measure presented in this chapter is where the delays are attributed. For existing measures, the delays are connected to the specific corridor where delay events occurred. However, the measure presented here is associated with the origin and destination of the trip. In essence, this measure connects the trips (and delays) to the zones that are economically affected by them instead of the zones where the delay physically occurs.
6.1.4 Chapter Organization

The remainder of this chapter is organized as follows. The following section on data processing describes the GPS data and the adopted approach used for this chapter. More detailed information on the processing approach can be found in Chapter 2 of the thesis. The algorithms were devised to process raw GPS data into trips conducted by individual trucks and calculate their travel time from the origin to destination. The next section of this chapter focuses on border delays for trips crossing the Canada-US border at several major crossing locations in Ontario. The international trips are examined to determine the delays occurring at the border (based on efforts conducted in Chapter 3) and the relative impact of this delay on the total trip.

This study is unique since the border delay is measured not just in absolute terms but also as a relative function of the full trip time. Following the border delay analysis, a section on expected and unexpected delays is discussed to differentiate between the types of delay that occur. This distinction is made due to the different impact that unexpected delays have on users. In addition, the chapter examines the spatial results of the proportional delay measure followed by an analysis of the role that distance exerts on delay. Finally, the Conclusions section provides a summary of the analysis and possibilities for future work based on the results.

6.2 Data Processing

A precursor to the analysis of truck movements in this paper is the conversion of raw GPS data into identifiable trips. A trip in this context represents a single leg journey observed of a truck from one location to another. However, this analysis does not explicitly account for combined trips – known as trip chaining - where trucks connect
multiple legs together. Similarly, truck tours are not analyzed where the truck returns to a start location after performing multiple stops (Hunt and Stefan, 2007).

The raw GPS data exists as a dataset of individual pings revealing the location (latitude/longitude) of a vehicle at a given point in time while also providing anonymous identifiers for the truck and corresponding carrier. While not included with the dataset, GPS based datasets may also provide: a dilution of precision (DOP) measure as an indication of spatial accuracy; the speed of the vehicle; engine characteristics if the GPS unit is connected to the vehicle’s electronic system; and the weight of the vehicle to determine full vehicles/empty backhauls and vehicle emissions. Moreover, some GPS systems utilize dead reckoning algorithms to impute the location of a truck if the GPS signal becomes too weak due to urban canyons or a minimal number of connected satellites. Connecting the individual GPS points for a given truck together sequentially based on the known time stamps provides a method of observing the movements of individual vehicles.

In this analysis, a dataset of GPS pings is utilized for the month of July, 2013. For this time period, the data corresponds to 30,000 individual trucks and 580 carriers. While the movements of the trucks occur across both Canada and the U.S, each of the carriers in the dataset is a Canadian owned company. Therefore the vehicle movements described in this chapter do not include the behaviour of U.S. based vehicles.

The detailed approach undertaken to identify trips is discussed earlier in Chapter 2. The major steps in this processing include an identification of stop events in the GPS dataset, a classification of stops as primary or secondary, and the creation of a dataset of
trips with the primary stop events used as endpoints. These three categories are discussed briefly in the next paragraphs.

Stop events are identified for vehicles that are observed within a 250 meter radius for 15 or longer. The 15 minute dwell time is larger than some studies focused on urban deliveries (such as the minimum 2 minute dwell time used by Hess et al., 2015) but considered acceptable in this application of inter-regional and international trips. The larger dwell time also helps remove the potential of false positives due to congestion. The busiest corridor of Highway 401 in Toronto, Ontario was analyzed to confirm that false positives due to congestion are not occurring (more information can be found in Section 2.2.2).

The observed stop events are further classified as primary or secondary stops. This classification is performed utilizing methods described in Chapter 2 including the entropy of carriers at a given location and the presence of a nearby firm identified as a gas station/rest stop. Primary stop events indicate that the truck stopped to transfer goods at the location. Secondary stop events are denoted when a truck is stopped for other purposes such as driver breaks or vehicle fueling. This is a necessary step since the raw GPS data does not provide any explicit indication of the reason for vehicle behaviors.

Since the transfer of goods is the goal of commercial truck movements, only primary stops are utilized as trip end points. In addition, a time constraint for valid trips is derived from dispatcher data on expected travel times between various regions. An expanded description of the process applied to obtain trips from the GPS pings is provided in Chapter 3. The initial trips for the single month of July total 221,800, providing a substantial sample of data to estimate delays. The truck trips traverse
locations between the west coast and east coast of North America and as far south as the U.S-Mexico border. However, the highest concentration of trips exists in Ontario. In particular, the Peel region observes the largest quantity of trips which is expected based on the high concentration of freight activity that occurs in that area. This pattern is expected due to the nature of the dataset pertaining to Canadian owned trucks.

The elapsed time observed for any trip includes the travel time for the vehicle as well as any time spent stopped for other purposes. The dwell time of stops was removed from the overall time since the duration of stops are not directly related to delay / congestion effects. Therefore the travel time for any given trip is calculated as follows:

\[ t_{i,\text{travel}} = t_{i,\text{elapsed}} - t_{i,\text{dwell}} \]  

(Eq. 6-1)

where \( t_{i,\text{travel}} \) is the total travel time pertaining to trip \( i \), \( t_{i,\text{elapsed}} \) is the measured elapsed time for the entire trip \( i \), and \( t_{i,\text{dwell}} \) is the dwell time belonging to any intermediate stops for trip \( i \).

### 6.3 Border Delays

The information derived from GPS data allows us to observe a vehicle for the entire duration of the trip. This affords a valuable opportunity to observe delays at a single point in time for a trip as well as its relative impact on the total trip. For example, cross-border delays can be measured as a proportion of the overall travel time. The delays associated with three crossings are analyzed in this chapter including – (1) Ambassador Bridge connecting Windsor ON, and Detroit MI (2) Blue Water Bridge connecting Sarnia ON and Port Huron, MI (3) Peace Bridge between Fort Erie, ON and Buffalo, NY. These three crossings are integral to the North American economy as the top three locations of
international crossing for commercial trips between Canada and the U.S. (MTO, 2016) as seen previously in Figure 1-3.

The crossing time for an international trip at the Canada-U.S. border is measured here by utilizing virtual perimeters (geofences) that surround the international crossing and the corresponding customs plazas for both countries. The geofences for the border crossings are provided in Figure 6-1. Since each trip is derived from individual pings, an interpolation of the crossing time is necessary to estimate when the truck crossed the geofence perimeter. Furthermore, the outer zone shown in Figure 6-1 ensures that the outside pings used in the interpolation are not located too far away from the site, where the accuracy of the interpolation results would be degraded. The crossing time estimated at one of the three border crossings pertains to the time necessary to make the crossing as well as the amount of time spent waiting at the customs inspection points located within the plaza.

Figure 6-1: Border geofences for three crossing locations
6.3.1 Border Crossing Trips and Spatial Patterns

In the month of July, 2013, the observed GPS based trip crossings in both directions at the Ambassador Bridge (AMB), Blue Water Bridge (BWB), and Peace Bridge (PCB) totaled 14,479, 6,187, and 8,387 respectively. While this forms a substantial dataset of truck crossing events, it is only a sample of the full population. According to the Public Border Operators Association (PBOA, 2017), the total truck crossings for the same month in 2013 included 175,258 (AMB), 125,182 (BWB), and 104,510 (PCB). A Map showing the origin of the GPS derived trips crossing the three border crossings is provided in Figure 6-2. This map includes a full year of data for 2013, but the pattern remains very similar from month to month.

Figure 6-2: Origin frequency of trips crossing three bridges (2013)
Past literature on these spatial patterns is sparse. However, a report from Parsons Brinckerhoff Quade & Douglas (PBQD, 2002) also provides a summary of truck based spatial flows across the Canada-U.S. border crossings based on the 1999 Canadian National Roadside Study. Their report, and the results in this chapter, both show the same 8 U.S. states as the main source of trips originating in the U.S. and crossing the Ambassador Bridge into Canada. These states include Michigan, Ohio, Illinois, Indiana, Texas, California, Tennessee, and Kentucky. However, it is worth noting that PBQD (2002) study suggests that Michigan makes up 42% of U.S. based origin trips crossing the Ambassador Bridge. By contrast, our study shows only 10% of the U.S. based trips arising from Michigan, with more trips spread out to other states. The discrepancy is likely partially due to the GPS dataset utilized here under-representing short trips (discussed in Chapter 4). In addition, the GPS dataset in this thesis is primarily composed of for-hire carriers, while lacking many large private fleets. In the automotive industry, this notably leaves out the major OEM automotive firms. Therefore a substantial portion of the automotive Just-In-Time (JIT) trips occurring between Michigan and Ontario are not included.

6.3.2 Border Crossing Travel Times

Statistics on the resulting border crossing times derived from GPS data are given in Table 6-1. The median border crossing time at the AMB, BWB, and PCB based on the observed crossing times are 13.6, 11.3, and 12.3 minutes respectively for vehicles headed to Canada. Moreover, the 95 percentile crossing time to measure the spread of crossing events is 42.3, 48.7, and 47.6 minutes for trucks headed to Canada. The 95 percentile crossing times as a measure of variability are tremendously important since a late arrival
of products caused by long delays can have a high cost associated with them. Moreover, since the cost of late shipments is typically much higher than the cost of an early shipment, carriers likely anticipate a crossing time much higher than the average (Anderson and Coates, 2010).

<table>
<thead>
<tr>
<th></th>
<th>Ambassador Bridge</th>
<th>Blue Water Bridge</th>
<th>Peace Bridge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAN bound</td>
<td>USA bound</td>
<td>CAN bound</td>
</tr>
<tr>
<td>Minimum</td>
<td>3.35</td>
<td>4.61</td>
<td>2.58</td>
</tr>
<tr>
<td>Median</td>
<td>13.6</td>
<td>14.3</td>
<td>11.3</td>
</tr>
<tr>
<td>Average</td>
<td>17.6</td>
<td>18.9</td>
<td>16.8</td>
</tr>
<tr>
<td>95 Perc.</td>
<td>42.3</td>
<td>48.3</td>
<td>48.7</td>
</tr>
<tr>
<td>Avg. trip delay proportion</td>
<td>2%</td>
<td>2.5%</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

For U.S. bound shipments, the crossing time tends to be higher compared to Canadian bound trips. This has been consistently the case in the various analyses of border crossing times conducted with the GPS dataset in this thesis. In particular, the Peace Bridge displays a large increase in the crossing time for trucks headed to the U.S. It is worth noting that the months of July and August are outliers for the Peace Bridge compared to the typical trend observed with the GPS data. For example, in 2013 and 2014, the July/August average crossing time is 26.4 minutes while all other months average 19.6 minutes. Therefore the larger crossing times for the Peace Bridge seen in Table 6-1 (and Figure 6-3 later in the chapter) are caused by the seasonal trend occurring in the summer months.
The delay at the border is calculated by identifying the minimum travel time necessary to cross the border (3.35 min AMB, 4.61 min BWB, and 2.58 min PCB) and subtracting this value from each crossing time. However, the observed border delays are only a portion of the typical travel time for a truck. For a short distance trip, the delay will comprise a relatively large portion of the total trip. Conversely, a long distance trip will typically observe a much smaller overall impact from border delays as a proportion of the total trip time. From the GPS dataset used in this chapter, the border crossing times are connected to the corresponding trips to provide information on the observed impact of delays on the entire trip. This process resulted in 10,316 trips that utilized the Ambassador Bridge, 4,001 trips that utilized the Blue Water Bridge, and 3,001 trips that utilized the Peace Bridge.

6.3.3 Proportion of Travel Time Affected by Border Delay

Statistics on the relative impact of a border delay are based on the travel time for a given trip and its corresponding border delay. For Canadian bound trips, the delay represents an average of 2% (AMB), 2.5% (BWB), and 3.2% (PCB) of the total trip travel time. Due to higher delays for U.S. bound trips, the average increases to 2.5% (AMB), 3.6% (BWB), and 4.4% (PCB). It is worth noting again that the Peace Bridge is higher in this case due to the seasonally larger border crossing times and not necessarily a reflection all months in the year.

A cross-section of the number of trips with a corresponding proportion of delay is shown in Figure 6-3. The majority of trips exhibit border delays that represent only 0-5% of the travel time. However, a noticeable quantity of trips exhibit border delays encompassing 5-10% of their travel time. A small proportion of trips experience delays
that account for more than 10% of the trip time. In reality, these extreme events are likely under-represented in the results for several reasons. First, delays greater than 90 minutes are removed from the dataset under the assumption that these trucks are undergoing secondary inspection, but this could also arise from extreme congestion. Moreover, during events where long delays occur, the queue of trucks can spill backwards onto the approach sections of the road that are not covered by the virtual perimeter (geofence) used to capture the border delays.

The large delays experienced relative to the total trip time can have a large impact on their costs. This is particularly true if their extensive delays lead to additional fees associated with late deliveries. Moreover, while many of the trips did not see substantial border delays, the potential for delays observed by a small number of trips likely causes most carriers to provide an extra buffer time (or hold extra inventory) that increases the indirect costs arising from uncertain border delays. While the actual costs of these delays cannot be easily estimated with the GPS data alone, Brown and Anderson (2015) estimate the extra ad valorem costs (as a percentage of the total value of goods) for carriers conducting cross-border trade between Canada and the U.S. varying on average between 0.4% to 0.9%, though this result only includes direct transportation costs.
Figure 6-3: Border delay as a proportion of the trip travel time for Canada bound (top) and U.S. bound (bottom) trips
6.4 Expected / Unexpected Trip Delays

In addition to border delays, vehicles encounter hindrances from various other sources such as commuter congestion on weekdays, construction activities, collisions, and extreme weather phenomenon. Delays caused by daily work commutes represent time that commuters expect to spend in congested traffic based on a given hour of the day with increasing delays during the morning and afternoon peak periods. While not commuters themselves, freight movements are similarly impacted by the temporal patterns of commuter traffic due to the shared nature of most major roads. Long-term construction projects also result in an expected additional travel time in most cases. By contrast, events such as inclement weather can be difficult to predict while traffic collisions are even more problematic. These types of events subsequently lead to unexpected delays.

Trips derived from GPS data in this analysis are organized by origin-destination paired zones. These zones are represented here by census divisions in Canada and MSA zones in the U.S. (along with counties filling in gaps between MSA zones). The time observed for a trip can be separated into three categories based on Figure 6-4. These three categories consist of (1) free flow travel (no delay), (2) free flow travel plus expected delays, or (3) free flow travel plus expected and unexpected delays. The first segment (left side) of the diagrams in Figure 6-4 represent the proportion of the trip where no delay occurs. In scenario 1, the trip only requires the minimum travel time to complete, therefore 100% of the trip is allocated to free flow travel (no delay). The minimum travel time was originally measured based on the observed minimum travel time identified for any trip between the origin and destination zones. However, the area and spatial
configuration of a zone erroneously biased the results (i.e. a trip to the near side of a zone will take less time than a trip to the far end of a zone). A revised method was introduced by calculating the maximum speed resulting from any trip for a given OD zone pair and utilizing this speed to determine the minimum possible travel time, effectively controlling for distance variation.

If a trip exceeds the minimum travel time, as seen in Scenario 2 (Figure 6-4), longer travel times above the minimum are labeled as expected delays. These delays are ‘expected’ as long as the travel time does not exceed the average travel time. An assumption included here is that travel times exceeding the minimum are not occurring due to variations in the free flow speed preferred by each driver. However, this is not as prominent of an issue for trucks compared to passenger vehicles. The average travel time was calculated based on two hour periods of the day (i.e. midnight to 2 AM, 2 AM to 4 AM, etc). For example, a vehicle traveling on a road segment at 4 PM would typically expect to encounter more commuter vehicles compared to the same location at 10 PM. For this second scenario, consider a trip where the actual travel time was 10 hours and the minimum travel time was 8 hours. The proportion of free flow travel would equal 80% while the proportion of expected delay would equal 20%.

Finally, if a trip exceeds the average travel time, as seen in Scenario 3 (Figure 6-4), we would observe some unexpected delays. In this scenario, consider a trip where the actual travel time was 11 hours, the average travel time was 10 hours, and the minimum travel time was 8 hours. In such a case, the proportion of free flow travel would equal 73% (8 hours /11 hours), the proportion of expected delay would equal 18% (2 hours /11 hours), and the proportion of unexpected delay would equal 9% (1 hour /11 hours).
The total number of trips from the one month GPS data resulted in 221,807 trips. However, only origin-destination (OD) pairs with at least 40 trips were utilized. This was necessary to obtain a suitable size of trips of any given origin-destination pair that can assess the three definitions of delay above. More specifically, the average travel time for a given time of day and the minimum travel time require enough data to establish a suitable baseline. The applied condition results in a smaller dataset of 83,654 trips.
belonging to 756 OD pairs. The average proportion of travel time from these trips associated with free flow travel, expected delays, and unexpected delays is 75%, 19%, and 6%, respectively. Therefore on average, 25% of the travel time for the trips in the GPS dataset was caused by some form of delay.

A map is provided in Figure 6-5 to show the average proportion of trip delays for the zones based on the origin and destination of trips. Therefore the map does not show where congestion occurs, but rather which zones produced or attracted trips that experienced delay. Overall, the average delay values for individual zones range from 10% to 34%. Two outlier zone pairs with average delay proportions above 50% were removed since they showed substantially inconsistent behavior.

Figure 6-5: Average trip delay proportion by zone
A breakdown of the trip patterns and delays for the top 6 zones from Figure 6-5 are provided in Figure 6-6 to help determine why these locations were affected by heavy delays. Among the top 6 zones, the number of connecting zones vary substantially. For example, zones in the western provinces of Alberta and British Columbia only have one and three connecting zones, respectively (Figure 6-6d and 6-6e). On the other hand, 32 zones are included for the Toronto zone shown in Figure 6-6f. Due to the limited number of connecting zones, the western zones in the top 6 may not be representative of all trips starting/ending in their respective areas. The small number of zones arises from the large restriction imposed on the number of trips needed between each origin-destination zone pair to correctly calibrate the average travel times by time of day. Relaxing this restriction could be performed to increase the number of zone pairs with delay statistics at the cost of a lower accuracy for the average times. Alternatively, if the minimum and average travel times could be generated from some other data/method, the restriction could be relaxed.
Figure 6-6c: Average trip delay for the Halton zone

Figure 6-6d: Average trip delay for Division No. 13 zone (Northwest of Edmonton)

Figure 6-6e: Average trip delay for the Fraser Valley zone

Figure 6-6f: Average trip delay for the Toronto zone
High volumes of traffic leading to congestion likely cause the large number of Toronto area zones to show up with the highest delays. In addition, the top 6 delay zones tend to have much shorter trips with 53% of trips occurring across less than 100 km compared to 17% for all other zones. The effect of distance may arise due to the smoothing out of specific delay events over a longer trip.

The average trip delays based on the distance travelled in 100 km intervals are shown in the solid line in Figure 6-7. The curve, starting from low to high distances (left to right), initially shows a decline in the average delays of trips as expected. However, a sharp rise occurs in the average delay for trips exhibiting distances greater than 900 km, thereby opposing the negative correlation expected between the delay and trip distance.

The inflection point of the original delays at 900 km is important due to the processing utilized to determine the dataset of valid trips. As discussed in Chapter 3, the GPS data processing includes a determination of viable travel times for trips. A relationship was derived from dispatcher information to determine the expected travel times of truck trips since trips with distances exceeding 900 km take a longer amount of time to complete due to the increased level of rest that is required (As shown in Figure 3-6 and Equation 3-1). Subsequently, the cut-off point for valid trips based on travel time is larger for the longer distance trips. The side-effect of the original time constraint is that some trips with extreme delay are included in the dataset that were not included for trips with distances less than 900 km, thereby causing an artificial rise in the curve at 900 km.

However, the dwell time from stops have been removed from the total trip time in this chapter (to more accurately reflect the actual travel time), negating the need for the separate time threshold originally placed on the trips traversing larger distances. As a
result, the time constraint imposed on trips to determine validity was adjusted to be consistent across all distances (i.e. the effective speed in Equation 3-1 was set to 70 km/hr for all trips) and resulted in the dotted curve in Figure 6-7. This adjusted pattern provides a better depiction of the relationship between delay and trip distance. Due to the consistent pattern observed using the adjusted dataset, the correlation between the two statistics (delay and distance) is a very strong -76.8% for the adjusted results but only -14.2% for the unadjusted dataset.

![Figure 6-7: Relationship between trip delay and distance](image)

### 6.5 Conclusions

This paper provides an overview of the delays observed for Canadian owned trucks. GPS data was processed to identify the movements of trucks along entire trips between zones instead of single locations. The trip duration was a key variable in the analysis, allowing the travel delays to be transformed into the proportion of each trip affected by delays on the road network. The analysis focused on two areas: (i) delays at the Canada-U.S border and (ii) expected/unexpected delays between origin/destination zonal pairs.
While delays at border crossings have been studied in the past, the measurement of delays as a function of travel time is considered unique for this chapter. The analysis of border delays was applied to three crossings that resulted in an average of 2% (AMB), 2.5% (BWB), and 3.2% (PCB) of the total trip travel time for Canadian bound trips. The larger proportion of delay attributed to the Peace Bridge is the result of shorter trip distances on average compared to the other two bridges since the crossing time statistics are relatively similar. On the U.S. side, the Peace Bridge also has a higher delay proportion but this is caused by the much higher crossing times during the summer period when the crossings in this chapter occurred.

The impact of delays on a full trip reflects more realistic importance on the costs involved for shippers compared to the singular effect of individual road segments. This type of measure will have more meaning in models that account for the costs of transportation. For example, economic models that can include a transportation linkage such as multi-regional input-output models and computable general equilibrium (CGE) models may benefit from an attribute measuring the proportion delays on the entire supply chain.

The overall delays for both domestic and international truck trips were deconstructed into the proportion of a trip with no delay (75% average), expected delay (19% average), and unexpected delay (6% average). The results were averaged for given origin and destination zones to provide a visual example of the results. Shorter trips tend to exhibit a larger delay proportion on average, with trips less than 200 km particularly noticeable with average delay proportions of 32.9% (0 to 100 km) and 25.5% (100 to 200 km).
While these delays are informative from the perspective of road users and policy makers, the measure of expected/unexpected delay may not be as useful for carriers in the supply chain process where a discrete amount of buffer time may be more commonly utilized (such as a 2 hour time window). A future focus on adapting this measure from the supply chain perspective may be a valuable exercise. Moreover, the locations exhibiting high average delays, such as those located in the Greater Toronto Area, can be further investigated to determine the specific road links where congestion is most problematic and therefore most urgently requiring attention. While the results in this chapter have pertained to measures of delay as a proportion of the whole trip, the GPS data can also be used to calculate delays at specific locations. However, local transportation agencies likely have existing data/technology in place to identify delays on busy roadways, making this functionality of GPS less appealing for this purpose. Regardless of where this information comes from, a two-step, top-down procedure could be implemented where the zones that are most impacted by delays are identified (as performed in this paper) first, followed by the identification of specific routes causing the delays. This could be an effective approach towards prioritizing infrastructure that require the most improvement.

6.6 Chapter 6 References


CHAPTER 7

CONCLUSIONS

7.1 Fulfilled Objectives

This dissertation addresses the challenge of adapting raw GPS data to extract knowledge regarding truck movement activities. Information on these truck patterns are currently lacking from the transportation literature, particularly knowledge pertaining to long distance trips and international (Canada-U.S.) freight trips. The creation of a dataset for these trip movements was completed using a source of opportunistic GPS data with over 1 billion GPS data pings per year that were originally created for the purpose of fleet management. The number of trips in the processed dataset includes an average of 245,000 trips per month, with data processed in this thesis for September 2012 to December 2014 and the month of March 2016. Information contained in the dataset of trips produced in this thesis did not previously exist at such a large volume and level of detail for Canadian long-haul truck trips and international crossings.

7.1.1 Data Processing

The successful processing of the raw GPS data into a set of trips required the implementation of several steps that were performed for this thesis. The first item was the creation of an overall processing approach to derive the dataset in Microsoft SQL. The latter platform was utilized due to its ability to handle (store and process) Big Data in a timely fashion. The details of the raw data characteristics and processing was the focus of Chapter 2, with a discussion of the overall implemented steps at the end of the chapter.
The spatial nature of transportation activities necessitate an examination of the truck data from a geographic perspective. ESRI ArcGIS software was relied upon heavily to test suspected hypotheses, but algorithms were then devised and adapted to Microsoft SQL to process the data records in a more efficient manner. For example, the entropy method discussed in detail in Chapter 2 is based on clustering (for the validation) or a buffer area (for the processing) that could be performed in GIS software (such as ArcGIS). Instead, these geo-processing methods were converted into scripts in Microsoft SQL (the entropy script is described as Script 6 at the end of Chapter 2). This type of work was instrumental towards properly processing the data. In the future, database software that is specifically designed for integrations with GIS functionality and allows for parallel processing may be beneficial towards improving processing efficiency compared to Microsoft SQL.

Finally, the nature of the processed GPS dataset as a sample of information required an analysis of the biases in order to understand the limitations inherited in the data. As a result, Chapter 4 of the thesis focused on several major biases discovered with respect to the processed GPS data. In addition, the chapter provided a novel method for adjusting the sample data appropriately in an expansion process. The objective was to provide a more realistic number of trips that match observed totals. The biases of the GPS data are discussed again in Section 6.3 discussing limitations in more detail.

7.1.2 Understanding the Impact of Border Crossings on Truck Freight

In addition to the data creation, this dissertation focused on applying the derived datasets towards an increased understanding of inter-regional and cross-border truck movements (first discussed in Section 1.7.2). Two pressing research questions were
posed in the thesis. The first question asks what factors influence the border crossing choice of trucks. This was addressed in Chapter 5 where a specific case study was used for trips between Toronto and Chicago due to the similar times that two international routes present (using the Ambassador Bridge and Blue Water Bridge).

The results from the analysis using a set of discrete choice models suggest that company/driver preference and the performance of the border crossing both play a role in the crossing choice decision of trucks. For example, some carriers were found to always select the same border crossing and trucks that previously used a border crossing were more likely to select the same location again. However, the average hourly crossing time was also found to be a determining factor in the decisions of trucks with a higher crossing time for a given hour deterring some vehicles and increasing the probability of utilizing the alternative border crossing. The crossing time at the Ambassador was tested in the model using the crossing time variable to determine the extent needed to create an even split instead of the observed 60/40 split favoring the Ambassador Bridge. Since a larger crossing time deters travelers, a sensitivity analysis was performed in which the variable was increased in increments until the 50/50 volume split was reached. The latter event occurred in the model results when the average hourly border crossing time at the Ambassador Bridge was increased by 50%.

The second posed question in the thesis asks how much of an effect border delay contributes towards the total time of an international trip. The processed trips in the dataset provide a unique perspective on this topic since delay at the border can be observed and further positioned with regards to the impact on the full trip. In response to the above research question, Chapter 3 of the thesis provides the final processing of the
GPS data into observable trips and provides a characterization of those trips with respect to several border crossings including the spatial patterns of the trips and crossing time trends. The observed trips from the GPS dataset that were derived in Chapter 3 are later analyzed in Chapter 6 with respect to delay.

The analysis in Chapter 6 estimated that the average amount of delay at the border crossings was generally small compared to the entire trip. The average proportion of delay at three major border crossings ranged from 2% to 3.2% for Canadian bound trucks and 2.5% to 4.4% for U.S. bound trucks. However, some trips experienced more substantial delays, including outliers greater than 20%. For Canadian bound trucks, a delay proportion in the range of 5% to 10% was experienced by 4% to 10% of observed trips (depending on crossing).

For U.S. bound trips, the same range of 5% to 10% delay proportions was experienced by 7% to 19% of trips observed in the dataset. For these trips, the observed delays at the border impose additional burdens to businesses that trade with the U.S. Alleviating this burden provides opportunities for economic growth. For example, a computable general equilibrium (CGE) model was used by Roberts et al. (2014) to show that adding an additional officer/inspection booth at a U.S. crossing (to reduce the crossing time into the U.S.) helped improve the Canadian export market and that in turn boosted U.S. economic growth.

7.2 Contributions of the Dissertation

7.2.1 Data creation

This dissertation addresses current gaps in transportation knowledge by creating a new dataset of truck trips from opportunistic GPS data and applying it to better
understand cross-border transportation behavior between Canada and the U.S. Traditionally, freight transportation has generally been considered an under-represented topic in past transportation research (Doustmohammadi et al., 2016). The lack of detailed data has been a major obstacle in the past. However, due to the large needs for freight data, new paradigms have emerged in recent years to tackle the issue of data scarcity. Chow et al. (2010) state that “..advances in applying data mining techniques to available or easily developed data sources would be a huge benefit to researchers and planners alike”.

The need for detailed data is particularly true for cross-border research, where the issue is compounded by a study area that spans multiple nations. As a result, the information produced as part of this thesis fills a major gap on Canadian truck movements across regions and the international Canada-U.S. border. This includes details such as trip origin/destination and border crossing. The large sample size of the processed data is also important since high survey costs tend to produce data that only spans a limited number of days or vehicles/carriers.

7.2.2 Data Mining Approach

The use of GPS data is gaining increased recognition for its value in transportation research. This thesis complements the work contained in other Canadian dissertations including the efforts conducted by Bryce Sharman at the University of Toronto with respect to data end point clustering and trip arrival rates for freight trips (Sharman, 2014), and the activity based processing efforts and map-matching for passenger GPS data produced by Ron Dalumpines at McMaster University (Dalumpines, 2014). This dissertation contributes to the adoption of GPS data in transportation by
providing an approach for converting raw GPS data into a viable source of inter-regional truck trips. Since similar GPS data to the one used in this research is likely to gain more attention in the future, the approaches devised in the thesis are highly transferable to future GPS applications. As such, the methods devised in this thesis (namely Chapters 2 and 3) can be used to process future GPS datasets that are likely to be in high demand.

One of the more practical contributions of the GPS processing presented in this thesis is the explicit separation of stop event purpose into two categories – primary and secondary. The primary stops become integral to the processing approach as they are used as endpoints for the identification of trips. As part of the stop purpose classification discussed in Chapter 2, this thesis also developed a new method of identifying secondary stops based on the characteristics of the GPS data itself. The method utilizes entropy to characterize the variety and distribution of carriers stopping at a given location, with high entropy being indicative of secondary stops. The patterns identified in this dissertation provide valuable insights into mining truck GPS data. These patterns allow an analyst to derive useful information from the raw data and create a richer processed dataset of trips. Moreover, the successful application of various patterns in the GPS data has broader implications towards the continued viability of opportunistic data for transportation research.

7.2.3 Characterization of Truck Trips

The characterization of the GPS derived truck trips in Chapter 3 presents a showcase of information that can be obtained at a macroscopic level (by aggregating trips) while Chapter 5 (border crossing choice) and Chapter 6 (trip delays) present applications of the data that derive information from individual trips. The usefulness of
GPS data as a source of aggregate and detailed information provides enormous opportunities in the future (discussed more in Section 7.3). Moreover, the spatial patterns of aggregate trips for specific border crossings (provided in Figures 3-8 to 3-17) provides greater insight into the relevance of the border crossing gateways to Canada and the U.S. For example, the patterns in the U.S. indicate that the Peace Bridge predominantly serves markets to the east of southwestern Ontario. By contrast, the Ambassador and Blue Water bridges tend to serve the western access to the U.S. Moreover, the two bridges are in close proximity to each other and therefore share some overlap in the pattern of origin/destination, but trips crossing the Ambassador bridge tend to spread out further (south towards Mexico and west towards the U.S. pacific coast).

In addition to spatial patterns, Chapter 3 also provides patterns on the crossing time distribution observed at the border crossings. This distribution is particularly important towards understanding the costs for supply chains to operate across the border, since the average crossing time is not necessarily as important as the variability (identified in Chapter 3 using the 90th percentile and 95th percentile crossing times). A large variability in crossing time requires firms to either anticipate a relatively long delay at the border or provide some other method of satisfying orders on the other side of the crossing to avoid incurring fines/work stoppages. Therefore the monitoring of delays is an important tool to monitor freight fluidity at the border. A valuable aspect of GPS data is that the processing can be repeated indefinitely over time so long as the original source of the data provides continuing access to GPS records. Therefore, any changes in the spatial patterns or crossing time distributions can be observed in the future.
7.2.4 Sample Expansion

A novel method of expanding GPS trip volumes was introduced in Chapter 4 of the thesis. The method performs two functions including (1) counteracting biases observed in the processed dataset of trips and (2) expanding the volume of trips to match observed totals. The expansion is an important consideration for future efforts that include a comprehensive model of freight patterns with an unbiased dataset that matches the full population of observed trips. As such, the expansion method put forward in this dissertation should be useful to transportation analysts and modellers.

7.2.5 Factors Influencing Border Crossing Choice

The derived GPS trips in the thesis are rich in terms of their spatial and temporal information. The derived trips were used in a case study to evaluate truck movement between Toronto and Chicago with a focus on border crossing choice behavior (in Chapter 5). The unique feature of this study arises from the similarity in travel times for the two alternative routes, especially that a typical route choice is predominantly based on minimizing travel times. As a result, the study asks what other types of factors can motivate the choice of crossing when the overall travel time is very similar.

The results from the analysis provide evidence of the importance that the crossing time at the border plays on the choice of crossing. This is particularly important when considering the impact that changes to one crossing (or the addition of a new crossing) can have on traffic demand. As such, the model will be of interest to policy analysts specialized in handling international cross border movements (e.g. Windsor-Detroit Bridge Authority), or transportation modellers who seek to estimate future demand.
7.2.6 Capturing the Influence of Trip Delays

The analysis presented in Chapter 6 quantifies the delay experienced by trucks travelling between regions in Canada and into the U.S. with an average delay equal to 25% of the total travel time. The chapter contributes towards a better understanding of delays encountered by trucks and further provides a new method for separating delay into expected and unexpected categories. The latter unexpected delays lead to additional costs associated with supply chain disruptions when moving goods by truck. As demonstrated in other contributions, this type of measure takes advantage of the GPS derived trips that observe the entire journey of a trip instead of one single location. Moreover, the analysis of delays can be utilized in a multitude of policy and modelling settings. For instance, on the policy side, the analysis can serve to provide a more realistic representation of the bottlenecks that curtail the movement of goods by truck over long distances. On the modeling side, the identified delays could be used to have a true representation of time in multi-regional economic models as opposed to using free-flow travel time.

7.3 Directions for Future Research

The utilized data in this dissertation pave the road for more work in the realm of freight transportation analysis using opportunistic GPS information. However, certain areas suffered some limitations that warrant further discussion. These limitations generally arise from the general characteristics of the data or the selected methodological approach employed to process the data. As such, future applications of the procedures presented in this thesis should acknowledge the key assumptions and limitations associated with these procedures. This is particularly important since the context of future applications may require adjustments or expansions to the current approaches. Also,
some of the procedures proposed in this thesis cannot be used in certain contexts. For instance, the allocation of secondary stop events to fuel stations for this thesis will ignore possible cases where the site is fueled as a primary purpose, therefore a study on fuel deliveries would not be suitable based on the approaches discussed in Chapter 2 of the thesis. The next sub-sections discuss the major limitations and provide direction towards areas of future research.

7.3.1 Inter-Regional versus urban truck movements

Both the data and processing utilized in this research were oriented towards long-haul trips. An analysis of the carriers owning the trucks generating the GPS data (as described in Section 2.3.2) suggests for-hire carriers who often take contracts for larger distances. By comparison, routine urban trips are more regularly conducted by private fleets. For example, the U.S. Commodity Flow Survey (CFS) methodology assumes that private fleet trucks travelling distances greater than 500 miles are actually for-hire vehicles based on the results of their model validations (USDOT, 2017).

The processing procedure offered in this research also favours long-haul truck movement since the thesis focuses on cross-border truck trips. For example, only trips that travel between Census Division zones (in Canada, as shown in Figure 3-2) are captured, resulting in inter-regional trips. Moreover, the processing of stop events was based on a threshold value of 15 minutes or longer. This dwell time threshold would be considered too long for shorter urban trips. However, this value was chosen due to the nature of long distance trips requiring longer dwell times at the stop location. The minimum dwell time set to 15 minutes is also large enough to avoid a noticeable occurrence of false positive stop events caused by heavy congestion (see Section 2.4).
The propensity of the resulting trips towards longer distances was confirmed in Chapter 4 when compared to data obtained from the 2006 commercial vehicle survey (CVS) conducted by the Ministry of Transportation Ontario (MTO).

As a result of the above preference towards long-haul trips, a study of urban goods movements would require an adjustment to several factors in the processing approach. First, the procedure will require a reduced minimum dwell time threshold to observe shorter goods deliveries. However, reducing the dwell time should be adjusted with caution to avoid false positive stop events. Second, smaller zones (e.g. census tracts) are more appropriate when dealing with urban goods movements. While zone size will not have much impact on the methodology presented in the thesis, data confidentiality may become an issue of concern to carriers.

7.3.2 Representation of Industry and Goods

In addition to the distance bias observed in the resulting dataset, Chapter 4 of the thesis also discussed a second bias related to industry representation. In particular, primary industries such as agriculture and mining are found to be severely under-represented. This issue is likely exacerbated by the method of industry classification utilized in the thesis. As discussed in Section 2.6, the industry associated with a stop event is determined by finding the nearest firm (within 200 meters). The dataset of firms are represented geographically as a single point location per firm. Subsequently, if the property/land parcel of a firm is large, a truck stopped at the property may not be within range of the listed point location of the firm.

In the future, a remedy for the above issue could be the utilization of property parcels instead of single point locations. As such, the parcels would be given the attribute
of the firm located on the property, and any stop events located on the property would likewise be associated with the property parcel and industry. Property parcels were utilized as part of the clustering approach for trip ends devised by Sharman and Roorda (2011). This approach was not followed in this thesis due to the difficulty in obtaining parcel property data covering our study area (i.e. the entirety of Canada and the U.S.). Moreover, this level of detail would require increased processing, which becomes troublesome for large datasets. Nevertheless, this approach could be utilized if the overall objective of the research work was particularly dependent on primary industries with larger properties. In addition to issues with primary industry discussed above, the lack of private carrier fleets in the GPS dataset indicate that industries favoring this type of transportation approach will be under-represented. For example, a large automotive producer such as Fiat Chrysler Automobiles (FCA) has its own private fleet to ship goods between facilities.

Finally, another limitation is that the industry is tied to the firms where the origin/destination of trips occurred, but is not a direct representation of the actual goods that were carried during the trip. This information would be difficult to obtain without direct driver input or a tracking system included in the supply chain process using technology such as radio frequency identification (RFID). Moreover, it is not known how much each truck is shipping by volume or weight (i.e. whether truck trip represents a full load, partial load, or empty-haul). Heavy trucks that are empty represent 15% of Canadian trucking activities based on vehicle-km (Transport Canada, 2011). In some GPS datasets, an estimation of vehicle weight may be possible if GPS pings in the dataset had very short time intervals (i.e. measured in seconds). In such a case, vehicle
acceleration/deceleration could be used as an indicator of empty trucks. Alternatively, empty trucks could be identified if the GPS pings are associated with fuel economy.

7.3.3 Truck Tours and Multi-modal Analysis

Passive GPS data could be used as a source to construct a complete truck tour model (Kuppam et al., 2014). Here, the movement of a truck is modeled as a round trip where the truck performs one or more stops (to transfer goods or take a break) before returning to the starting point. The raw GPS data used in this thesis has already been used to create several components of a reliable truck tour model in Gingerich and Maoh (2015). These components included a stop frequency model to determine the number of stops for a given tour and a stop purpose model to determine the purpose for each stop (whether it is secondary, primary, or a return to base).

Since the GPS data in this thesis does not track the movement of individual goods, explicit multi-modal analysis is not possible. However, the data can still serve as a supplementary source to study the impacts of multi-modal infrastructure on truck movement patterns. For example, an analysis was conducted using the GPS derived trips to determine the inter-relation between airport proximity and warehousing based transportation (Gingerich and Maoh, 2017). The results indicate that warehousing truck trips (by origin or destination) near a major airport travel 1.8 times further than other warehousing trips. These airport-based truck trips also provide evidence of heterogeneous patterns of goods movement due to multi-modal activities. Future research in this area could focus on the role of other types of inter-modal facilities (namely: rail yards and marine ports) on the patterns of truck trips.
7.3.4 Data Integration

The derived information from the processed GPS records in this thesis was instrumental to understand regional and cross-border truck movement patterns. Besides being an excellent source for freight transportation modeling (as illustrated in Chapter 5 and Chapter 6), the obtained trips can be used to assist federal agencies such as Transport Canada and Statistics Canada to add value to their current datasets. For instance, data from the Canadian Vehicle Use Survey, which is collected and maintained by Transport Canada (Leore, 2015), can be expanded using the GPS trips used in this thesis. The CVUS data is based on a smaller sample of trucks compared to the Shaw Tracking data. However, the CVUS GPS records contain pings at 1 second intervals and provides active data logging when a stop is performed to describe the nature of the stop. As such, future work could focus on integrating both sources via various data fusion techniques. The key point here is that any source of GPS truck data which provides a smaller sample with more detail can be a beneficial complement to the Shaw Tracking GPS data used here (where a much larger sample exists but fewer details are available).

In addition to the benefits of a second source of GPS data, stated preference (hypothetical) surveys could be useful additions to this research to expand on some of the revealed preference (observed) results. For example, the results of the border crossing choice revealed that some carriers will consistently use one border crossing. These results could be validated and further investigated using stated preference surveys to identify why some carriers are observed to only use one crossing (i.e. is it due to familiarity, the nearby presence of a cross-border shipping service, and so on and so forth) and what would be required to cause their decision to change. As such, stated preference surveys
can provide more complementary information that will enrich the GPS data used in this thesis.

7.3.5 Route Choice Modelling

Chapter 5 in the thesis presented the results from studying the border crossing choice decision made by trucks. There is still far greater potential of using the GPS dataset to identify the routes chosen by trucks to move goods between regions (and countries). GPS pings (point based data) can be converted to routes along road segments (line based data) using a process called map-matching (see for example Dalumpines, 2014; Dhakar, 2014). Individual trips from the map-matching process can be utilized for route choice modelling. These trips are particularly well suited for addressing route choice problems since one can observe the full behavior of trips from start to finish. The individual map-matched trips can be further aggregated to determine the volume of trucks on road links, as shown in Figure 7-1. The latter map is based on map-matched trips for a one week dataset of GPS pings that were generated in March, 2016. The generated information can then be used to identify the most critical links in the network.
7.4 Final Remarks

In recent years, an enormous volume of information amounting to 2.5 quintillion ($10^{18}$) bytes is generated on a daily basis. According to IBM (2013), 90% of the digital data that existed in 2013 was created in the previous 2 years. The surge of digital information from various sources, including GPS transponders, is reshaping the landscape of many disciplines including freight transportation. Tremendous new opportunities are likely to emerge in the future as the amount of information created continues to increase. The availability of data generated from information and telecommunication sources such as GPS, bluetooth, cell phones and social media will be
a game changer for transportation researchers. Moreover, as connected and autonomous vehicles become a modern reality, more detailed data are likely to become available for analysis. Connected/Autonomous vehicles essentially transform from traditional vehicles into mobile data probes with up to 30 terabytes (TB) of data generated each day per vehicle (SAS, 2015).

The availability of large volumes of opportunistic GPS data presents cost-effective solutions to leverage existing/traditional data. However, the drawback to the former type of data source is the processing required to convert the raw information into a suitable source for transportation research. This thesis therefore provides seminal efforts for advancing the state of transportation practice. It does so by demonstrating that passive GPS truck data are capable of providing new insights into the patterns of transportation movements between regions and the busiest freight crossing gateways between Canada and the U.S.

7.5 Chapter 7 References

IBM (2013) 2.5 quintillion bytes of data created every day. How does CPG & Retail manage it? [Online blog] <https://www.ibm.com/blogs/insights-on-business/consumer-
products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/>.


APPENDIX A

SQL SCRIPTS

The following scripts are written in the SQL language and implemented in Microsoft SQL Server 2008 R2 software. These scripts provide an example of the structure used to process the data, but caution should be used when adapting them for other analyses. The scripts follow the order outlined previously in Chapter 2, Figure 2-15.

Script 1: Creation of Tables

Purpose: Initializes a series of tables that will be populated with data in future scripts.

<Script 1 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

CREATE TABLE [dbo].[ProcessedData_Oct1_2013] ---User Input (Processed GPS Table)
  ([SeqID] [bigint] NULL,
   [SeqPID] [bigint] NULL,
   [CID] [varchar](25) NULL,
   [PID] [varchar](25) NULL,
   [CPID] [varchar](50) NULL,
   [Latitude] [real] NULL,
   [Longitude] [real] NULL,
   [TimeEntry] [datetime] NULL,
   [ElapsedTime] [datetime] NULL,
   [DwellTime] [datetime] NULL,
   [Alpha] [decimal](18, 5) NULL,
   [Stop] [int] NULL,
   [SeqID_Cleaned] [bigint] NULL,
   [TourID] [int] NULL
  ) ON [PRIMARY]
GO

CREATE TABLE [dbo].[DwellTime](
  [SeqID] [bigint] NULL,
  [DwellTime] [datetime] NULL
  ) ON [PRIMARY]
GO

<Script 1 End>
CREATE TABLE [dbo].[Trajectory](
    [SeqID] [bigint] NULL,
    [Alpha] [decimal](18, 5) NULL
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[OriginLatitude](
    [Carrier] [varchar](10) NULL,
    [AvgLatitude] [decimal](18, 5) NULL
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[OriginLongitude](
    [Carrier] [varchar](10) NULL,
    [AvgLongitude] [decimal](18, 5) NULL
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[ShippingDepots](
    [carrier] [varchar](10) NULL,
    [Latitude] [decimal](18, 5) NULL,
    [Longitude] [decimal](18, 5) NULL
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[First_Record](
    [SeqID] [bigint] NULL,
    [SeqPID] [bigint] NULL,
    [CID] [varchar](25) NULL,
    [PID] [varchar](25) NULL,
    [CPID] [varchar](50) NULL,
    [Latitude] [real] NULL,
    [Longitude] [real] NULL,
    [TimeEntry] [datetime] NULL,
    [ElapsedTime] [datetime] NULL,
    [DwellTime] [datetime] NULL,
    [Alpha] [decimal](18, 5) NULL,
    [Stop] [int] NULL,
    [SeqID_Cleaned] [bigint] NULL,
    [TourID] [int] NULL,
    [Combo_coordinate] [decimal](18, 5) NULL
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[Stops](
    [SeqID] [bigint] NULL,
    [SeqID_Cleaned] [bigint] NULL,
    [CID] [int] NULL,
    [PID] [bigint] NULL,
    [CPID] [varchar] (50) NULL,
    [Latitude] [real] NULL,
    [Longitude] [real] NULL,
    [ComboCoordinate] [decimal](18, 0) NULL,
[TimeEntry] [datetime] NULL,
[DwellHours] [decimal](18, 5) NULL,
[TourStart] [int] NULL,
[TourID] [bigint] NULL,
[TotalTrucks] [int] NULL,
[TotalCarriers] [int] NULL,
[EntropyIndex] [decimal](5, 2) NULL,
[HeadingChange] [decimal](10, 5) NULL,
[HeadingPre5] [Decimal](18, 5) NULL,
[HeadingPost5] [Decimal](18, 5)
) ON [PRIMARY]
GO

<Script 1 End>
Script 2: Processing Raw GPS Data Pings

Purpose: Transfers raw data into a new table and begins initial processing.

<Script 2 Begin>

/* INSERT ORIGINAL DATA INTO A NEW TABLE */
USE [Temp_OneDay_Test] ---User Input (Database)
GO

/* User Input: Field mapping for CID, PID, Latitude, Longitude, and TimeEntry */
;with CTE as
  (SELECT
   [Column 0] AS [CID],
   [Column 1] AS [PID],
   CAST([Column 0] as varchar(25)) + '-' + CAST([Column 1] as varchar(25)) as CPID,
   [Column 2] AS [Latitude],
   [Column 3] AS [Longitude],
   CONVERT(DATETIME,STUFF(STUFF(STUFF(CAST([Column 4] as varchar(8)) + [Column 5],9,0,' '),12,0,':'),15,0,':')) as TimeEntry,
   NULL as [ElapsedTime],
   NULL as [DwellTime],
   NULL as [Alpha],
   NULL as [Stop],
   NULL as [SeqID_Cleaned],
   NULL as [TourID]
   FROM [dbo].[Oct1_ShawData]) ---User Input (Raw GPS Table)

INSERT INTO [dbo].[ProcessedData_Oct1_2013] ---User Input (Processed GPS Table)
SELECT
   ROW_NUMBER() OVER (ORDER BY [CPID],[TimeEntry]) AS [SeqID],
   ROW_NUMBER() OVER (PARTITION BY [CPID] ORDER BY [TimeEntry]) - 1
   AS [SeqPID],
   [CID],
   [PID],
   [CPID],
   [Latitude],
   [Longitude],
   [TimeEntry],
   [ElapsedTime],
   [DwellTime],
   [Alpha],
   [Stop],
   [SeqID_Cleaned],
   [TourID]
FROM CTE

/* CALCULATE THE ELAPSED TIME BETWEEN POINTS */
;WITH CTE AS
  (SELECT a.SeqID AS [SeqID],
   a.ElapsedTime AS [ElapsedTime]
, a.TimeEntry - b.TimeEntry AS [TempElapsedTime]
FROM [dbo].[ProcessedData_Oct1_2013] a ---User Input (Processed GPS Table)
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] b ---User Input (Processed GPS Table)
ON a.SeqID - 1 = b.SeqID
)

UPDATE CTE
SET [ElapsedTime] = [TempElapsedTime]
GO

Update [dbo].[ProcessedData_Oct1_2013] ---User Input (Processed GPS Table)
Set [ElapsedTime] = NULL Where [SeqPID] = 0
GO

/* CALCULATE DWELL TIME */
Declare @latitude as real, @longitude as real, @timeentry as datetime,
@SeqID as bigint, @SeqPID as bigint, @elapsedtime as datetime,
@DwellTime as Datetime, @LatStart as Decimal(18,5), @LongStart as Decimal (18,5),
@LatEnd as Decimal (18,5), @LongEnd as Decimal (18,5),
@Distance as decimal (18,5),
@CID as Varchar(25),@PID as Varchar(25),@CPID as Varchar(50),@Alpha as Decimal (18,5),@Stop as int, @SeqID_Cleaned as bigint, @TourID as int

/*Set cursor to input table*/
declare db_cursor CURSOR FOR
Select
[SeqID],[SeqPID],[CID],[PID],[CPID],[Latitude],[Longitude],[TimeEntry],[ElapsedTime]
From [dbo].[ProcessedData_Oct1_2013] ---User Input (Processed GPS Table)
Order by [SeqID]

/*Open cursor and fetch the first record*/
Open db_cursor
Fetch NEXT From db_cursor into
@SeqID,@SeqPID,@CID,@PID,@CPID,@latitude, @longitude, @timeentry,
@elapsedtime

WHILE (@@FETCH_STATUS = 0)
BEGIN

/*Set the current coordinates as the latitude and longitude end coordinates*/
Set @LatEnd = @latitude
Set @LongEnd = @longitude
Set @Distance = SQRT((@LatEnd - @LatStart)*(@LatEnd - @LatStart) + (@LongEnd - @LongStart)*(@LongEnd - @LongStart))

/*Set the dwelltime equal to null for the first record of each truck*/
IF (@SeqPID = 0)
Begin
Insert into [dbo].[DwellTime]
DECLARING VARIABLES

```sql
([SeqID],[DwellTime])
```

```
Values (@SeqID, NULL)
Set @LatStart = @latitude
Set @LongStart = @longitude
Set @DwellTime = 0
Set @Distance = NULL
```

```
Fetch NEXT From db_cursor into
@SeqID, @SeqPID, @CID, @PID, @CPID, @latitude, @longitude, @timeentry,
@elapsedtime
```

```sql
End
```

```
Else
/*If the vehicle moved a significant distance, the dwell
time is set to 0
and the new start coordinates are set as the current
coordinates before fetching the next record*/
IF(@Distance > 0.0025)
Begin
    Insert into [dbo].[DwellTime]
    ([SeqID],[DwellTime])
    Values (@SeqID, 0)
    Set @LatStart = @latitude
    Set @LongStart = @longitude
    Set @DwellTime = 0
    Fetch NEXT From db_cursor into
@SeqID, @SeqPID, @CID, @PID, @CPID, @latitude, @longitude, @timeentry,
@elapsedtime
End
/*If the vehicle did not move, the dwell time is
cumulatively added to the previous dwell time and
added to the new table before fetching the next record*/
IF(@Distance <= 0.0025)
Begin
    Set @DwellTime = (@DwellTime + @elapsedtime)
    Insert into [dbo].[DwellTime]
    ([SeqID],[DwellTime])
    Values (@SeqID, @DwellTime)
    Fetch NEXT From db_cursor into
@SeqID, @SeqPID, @CID, @PID, @CPID, @latitude, @longitude, @timeentry,
@elapsedtime
End
```

Close db_cursor
Deallocate db_cursor
GO

```
UPDATE a
SET [DwellTime] = b.DwellTime
FROM [dbo].[ProcessedData_Oct1_2013] a
```
(SELECT a.SeqID as SeqID,
  ,a.SeqPID as SeqPID,
  ,a.CID as CID,
  ,a.PID as PID,
  ,a.CPID as CPID,
  ,a.Latitude as Latitude,
  ,a.Longitude as Longitude,
  ,a.TimeEntry as TimeEntry,
  ,a.ElapsedTime as ElapsedTime,
  ,a.DwellTime as DwellTime,
  ,a.Latitude - b.Latitude as Lat21,
  ,a.Longitude - b.Longitude as Long21
  ,Theta = CASE (a.Latitude - b.Latitude)
      When 0 THEN CASE (SIGN(a.Longitude - b.Longitude))
        When -1 THEN 270
        When 1 THEN 90
        When 0 THEN NULL
      END
    Else ATAN((a.Longitude - b.Longitude) / (a.Latitude - b.Latitude)) * 180/PI()
    END
FROM [dbo].[ProcessedData_Oct1_2013] a
  LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] b
  ON a.SeqID - 1 = b.SeqID )

INSERT INTO [dbo].[Trajectory]
SELECT SeqID,
  ,Alpha = CASE Sign(Lat21)
        When -1 THEN (180 + Theta)
        When 1 THEN CASE SIGN(Long21)
          When -1 THEN (360 + Theta)
          When 1 THEN (Theta)
          When 0 THEN 0
          END
        When 0 THEN (Theta)
        END
FROM CTE
END

UPDATE a
SET [Alpha] = b.Alpha
FROM [dbo].[ProcessedData_Oct1_2013] a
  LEFT OUTER JOIN [dbo].[Trajectory] b
  ON a.SeqID = b.SeqID
GO

UPDATE [dbo].[ProcessedData_Oct1_2013]
SET [Alpha] = NULL
WHERE [SeqPID] = '0'
GO

/* CALCULATE Stops WITH Dwell Times GREATER THAN 15 MINUTES */
WITH CTE AS

(SELECT a.[SeqID] as SeqID,
    a.[DwellTime] as DwellTime,
    b.[DwellTime] as DwellTimeLead,
    a.[Stop] as [Stop],
    [StopTemp] = CASE(SIGN(CAST(a.[DwellTime] as Decimal(18,5))*24 - 0.25))
        When 1 THEN CASE(ISNULL(b.[DwellTime],0))
        When 0 THEN 1
        Else NULL
        END
        Else NULL
    END
FROM [dbo].[ProcessedData_Oct1_2013] a  ---User Input (Processed GPS Table)
    LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] b  ---User Input (Processed GPS Table)
    ON a.SeqID + 1 = b.SeqID
)

UPDATE CTE
SET [Stop] = [StopTemp]
GO

/* REMOVE (CLEAN) INTERMEDIATE POINTS BETWEEN A STOP START/END */
DELETE a
FROM [dbo].[ProcessedData_Oct1_2013] a  ---User Input (Processed GPS Table)
    LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] b  ---User Input (Processed GPS Table)
    ON a.SeqID + 1 = b.SeqID
WHERE CAST(a.[DwellTime] as Decimal(18,5))*24 > 0 and
    CAST(b.[DwellTime] as Decimal(18,5))*24 > 0
GO

/* CALCULATE A NEW SEQUENTIAL ID BASED ON THE PREVIOUSLY CLEANED DATA RECORDS */
WITH CTE AS

(SELECT [SeqID],
    [SeqPID],
    [CID],
    [PID],
    [CPID],
    [Latitude],
    [Longitude],
    [TimeEntry],
    [ElapsedTime],
    [DwellTime],
    [Alpha],
    [Stop],
    [SeqID_Cleaned],
    [TourID],
    ROW_NUMBER() OVER (ORDER BY SeqID) as RN
FROM [dbo].[ProcessedData_Oct1_2013]  ---User Input (Processed GPS Table)
)
UPDATE CTE
SET [SeqID_Cleaned] = [RN]
GO

/* CREATE INDEXES ON SEVERAL VARIABLES IN THE MAIN GPS PING TABLE */
CREATE INDEX ix_ProcessedData_Oct1_2013_SeqID
ON [dbo].[ProcessedData_Oct1_2013]([SeqID])  ---User Input (Processed GPS Table)
GO

CREATE INDEX ix_ProcessedData_Oct1_2013_SeqID_Desc
ON [dbo].[ProcessedData_Oct1_2013]([SeqID] DESC)  ---User Input (Processed GPS Table)
GO

CREATE INDEX ix_ProcessedData_Oct1_2013_TimeEntry
ON [dbo].[ProcessedData_Oct1_2013]([TimeEntry])  ---User Input (Processed GPS Table)
GO

CREATE INDEX ix_ProcessedData_Oct1_2013_Lat_Long
ON [dbo].[ProcessedData_Oct1_2013]([Latitude], [Longitude])  ---User Input (Processed GPS Table)
GO

CREATE INDEX ix_ProcessedData_Oct1_2013_CPID
ON [dbo].[ProcessedData_Oct1_2013]([CPID])  ---User Input (Processed GPS Table)
GO

CREATE INDEX ix_ProcessedData_Oct1_2013_SeqID_Cleaned
ON [dbo].[ProcessedData_Oct1_2013]([SeqID_Cleaned])  ---User Input (Processed GPS Table)
GO

<Script 2 End>
Script 3: Computing Carrier Statistics

Script Purpose: Provides a summary of carrier statistics.

<Script 3 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

SET ANSI_NULLS ON
GO

SET QUOTED_IDENTIFIER ON
GO

CREATE TABLE [dbo].[CarrierSummary](
    [CID] [smallint] NULL,
    [TruckCount] [int] NULL,
    [RowNumber] [int] NULL)
ON [PRIMARY]
GO

DELETE FROM [dbo].[CarrierSummary]
GO

INSERT INTO [dbo].[CarrierSummary]
SELECT [CID], COUNT(DISTINCT [CPID]) as TruckCount,
       ROW_NUMBER() OVER(ORDER BY [CID]) AS RowNumber
FROM [dbo].[ProcessedData_Oct1_2013] ---User Input (Processed GPS Table)
GROUP BY [CID]
ORDER BY [CID]
GO

<Script 3 End>
Script 4: Identifying Carrier Shipping Depots

Purpose: Identifies one location per carrier pertaining to a shipping depot.

<Script 4 Begin>

/* This script locates the origin shipping yards for each carrier in the dataset */

/** The Carrier Summary table should be up to date (using Script 3) ***/

USE [GPS_Year2_2014] ---User Input (Database)
GO

Declare @carrier as varchar(8), @counter as int, @TotalCarriers as int
Set @counter = 1
Set @TotalCarriers = (SELECT MAX([RowNumber])as MaxRow FROM [dbo].[CarrierSummary])

While (@counter <= @TotalCarriers)
Begin
    Set @carrier = (SELECT [CID]
FROM [dbo].[CarrierSummary] WHERE [RowNumber] = @counter)
    Print @carrier
    If
    (SELECT TOP 1 Count(CAST([Latitude] as decimal(18,2)))
    FROM [dbo].[First_Record]
    WHERE [CID] = @carrier and [Combo_coordinate] =
    (SELECT TOP 1 [Combo_coordinate]
    FROM [dbo].[First_Record]
    WHERE [CID] = @carrier
    GROUP BY [Combo_coordinate]
    ORDER BY COUNT(*) DESC)
    < 3
    Begin
        insert into [dbo].[OriginLatitude]
        SELECT @carrier as carrier, null as AvgLatitude
    End
    Else
        insert into [dbo].[OriginLatitude]
        SELECT @carrier as carrier,
        ,CAST(AVG([Latitude])as decimal(18,5)) as AvgLatitude
    End

/*Take the average latitude (to 5 significant digits) from records selected below*/
FROM [dbo].[First_Record]
WHERE [Latitude] In

/*Select Records with latitude corresponding to the latitude and combo mode found below*/
(SELECT [Latitude]
FROM [dbo].[First_Record]
WHERE [CID] = @carrier and CAST([Latitude] as decimal(18,2)) =
(SELECT TOP 1 CAST([Latitude] as decimal(18,2))
/*Calculate mode of latitude values that corresponds to the combo mode*/
FROM [dbo].[First_Record]
WHERE [CID] = @carrier and [Combo_coordinate] =
(SELECT TOP 1 [Combo_coordinate]
FROM [dbo].[First_Record]
WHERE [CID] = @carrier
GROUP BY [Combo_coordinate]
ORDER BY COUNT(*) DESC)
GROUP BY CAST([Latitude] as decimal(18,2))
ORDER BY COUNT([Latitude] as decimal(18,2)) DESC)
and [Combo_coordinate] =
(SELECT TOP 1 [Combo_coordinate]
FROM [dbo].[First_Record]
WHERE [CID] = @carrier
GROUP BY [Combo_coordinate]
ORDER BY COUNT(*) DESC)
/*Calculate Longitude Coordinate*/
If
(SELECT TOP 1 Count(CAST([Longitude] as decimal(18,2)))
FROM [dbo].[First_Record]
WHERE [CID] = @carrier and [Combo_coordinate] =
(SELECT TOP 1 [Combo_coordinate]
FROM [dbo].[First_Record]
WHERE [CID] = @carrier
GROUP BY [Combo_coordinate]
ORDER BY COUNT(*) DESC)
GROUP BY CAST([Longitude] as decimal(18,2))
ORDER BY COUNT(CAST([Longitude] as decimal(18,2))) DESC) < 3
Begin
insert into [dbo].[OriginLongitude]
SELECT @carrier as carrier, null as AvgLongitude
End
Else
insert into [dbo].[OriginLongitude]
SELECT @carrier as carrier,
CAST(AVG([Longitude])as decimal(18,5)) as AvgLongitude
FROM [dbo].[First_Record]
WHERE [Longitude] in
(SELECT [Longitude]
FROM [dbo].[First_Record]
WHERE [CID] = @carrier and CAST([Longitude] as decimal(18,2)) =
(SELECT TOP 1 CAST([Longitude] as decimal(18,2))
FROM [dbo].[First_Record]
WHERE [CID] = @carrier and [Combo_coordinate] =
(SELECT TOP 1 [Combo_coordinate]
FROM [dbo].[First_Record]
WHERE [CID] = @carrier
GROUP BY [Combo_coordinate]
ORDER BY COUNT(*) DESC)
/*Combine latitude and longitude origin values into one table*/
Insert Into [dbo].[ShippingDepots]
SELECT [dbo].[OriginLatitude].carrier,
[dbo].[OriginLatitude].AvgLatitude,[dbo].[OriginLongitude].AvgLongitude
FROM [dbo].[OriginLatitude]
INNER JOIN [dbo].[OriginLongitude]
ON [dbo].[OriginLatitude].carrier = [dbo].[OriginLongitude].carrier
GO

<Script 4 End>

Script 5: Creating a Dataset of Stop Events

Purpose: Creates a dataset exclusively created to hold stop events and adds some information such as heading change before/after stop (not used/discussed in the thesis).

<Script 5 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

INSERT INTO [dbo].[Stops]
SELECT [SeqID]
,[SeqID_Cleaned]
,[CID]
,[PID]
,[CPID]
,[Latitude]
,[Longitude]
,CAST(([Latitude] * [Longitude]) as Decimal(18,0)) as [ComboCoordinate]
,[TimeEntry]
,CAST([DwellTime] as decimal(18,5)) * 24 AS [DwellHours]
,NULL as [TourStart]
,NULL as [TourID]
,NULL as [TotalTrucks]
FROM [dbo].[ProcessedData_Oct1_2013]  ---User Input (Processed GPS Table)
WHERE [Stop] = 1
GO

CREATE INDEX ix_Stops_SeqID_Cleaned
ON [dbo].[Stops] ([SeqID_Cleaned])
GO

;WITH CTE AS
(SELECT TOP 1000000000
  a.[Latitude],
  a.[Longitude],
  a.[TimeEntry],
  a.[SeqID_Cleaned],
  a.[HeadingPre5],
  b.[Alpha] AS AlphaPre1,
  c.[Alpha] AS AlphaPre2,
  d.[Alpha] AS AlphaPre3,
  e.[Alpha] AS AlphaPre4,
  f.[Alpha] AS AlphaPre5,
  b.[Stop] AS StopPre1,
  c.[Stop] AS StopPre2,
  d.[Stop] AS StopPre3,
  e.[Stop] AS StopPre4,
  f.[Stop] AS StopPre5
FROM [dbo].[Stops] a
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] b  ---User Input (Processed GPS Table)
ON a.[SeqID_Cleaned] - 1 = b.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] c  ---User Input (Processed GPS Table)
ON a.[SeqID_Cleaned] - 2 = c.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] d  ---User Input (Processed GPS Table)
ON a.[SeqID_Cleaned] - 3 = d.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] e  ---User Input (Processed GPS Table)
ON a.[SeqID_Cleaned] - 4 = e.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] f  ---User Input (Processed GPS Table)
ON a.[SeqID_Cleaned] - 5 = f.[SeqID_Cleaned]
)

UPDATE CTE SET [HeadingPre5] =
(CASE WHEN AlphaPre1 is null or StopPre1 = 1 THEN null ELSE
  (CASE WHEN AlphaPre2 is null or StopPre2 = 1 THEN AlphaPre1 ELSE
    (CASE WHEN AlphaPre3 is null or StopPre3 = 1 THEN (AlphaPre1 + AlphaPre2) / 2 ELSE
      (CASE WHEN AlphaPre4 is null or StopPre4 = 1 THEN (AlphaPre1 + AlphaPre2 + AlphaPre3) / 3 ELSE
        (CASE WHEN AlphaPre5 is null or StopPre5 = 1 THEN (AlphaPre1 + AlphaPre2 + AlphaPre3 + AlphaPre4) / 4 ELSE
          (CASE WHEN AlphaPre1 is not null THEN AlphaPre1 ELSE
            (CASE WHEN AlphaPre2 is not null THEN AlphaPre2 ELSE
              (CASE WHEN AlphaPre3 is not null THEN AlphaPre3 ELSE
                (CASE WHEN AlphaPre4 is not null THEN AlphaPre4 ELSE
                  (CASE WHEN AlphaPre5 is not null THEN AlphaPre5 ELSE
                    null END) END) END) END) END) END) END) END) END) END) END)
(CASE When AlphaPre4 is null or StopPre4 = 1 THEN (AlphaPre1 + AlphaPre2 + AlphaPre3) / 3 ELSE (CASE When AlphaPre5 is null or StopPre5 = 1 THEN (AlphaPre1 + AlphaPre2 + AlphaPre3 + AlphaPre4) / 4 ELSE (AlphaPre1 + AlphaPre2 + AlphaPre3 + AlphaPre4 + AlphaPre5) / 5 END) END) END) END) END)

From CTE

GO

; With CTE as

(SELECT TOP 1000000000

   a.[Latitude],
   a.[Longitude],
   a.[TimeEntry],
   a.[SeqID_Cleaned],
   a.[HeadingPost5],
   b.[Alpha] as AlphaPost1,
   c.[Alpha] as AlphaPost2,
   d.[Alpha] as AlphaPost3,
   e.[Alpha] as AlphaPost4,
   f.[Alpha] as AlphaPost5,
   b.[Stop] as StopPost1,
   c.[Stop] as StopPost2,
   d.[Stop] as StopPost3,
   e.[Stop] as StopPost4,
   f.[Stop] as StopPost5

FROM [dbo].[Stops] a
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] b
   ---User Input
   (Processed GPS Table)
   ON a.[SeqID_Cleaned] + 1 = b.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] c
   ---User Input
   (Processed GPS Table)
   ON a.[SeqID_Cleaned] + 2 = c.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] d
   ---User Input
   (Processed GPS Table)
   ON a.[SeqID_Cleaned] + 3 = d.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] e
   ---User Input
   (Processed GPS Table)
   ON a.[SeqID_Cleaned] + 4 = e.[SeqID_Cleaned]
LEFT OUTER JOIN [dbo].[ProcessedData_Oct1_2013] f
   ---User Input
   (Processed GPS Table)
   ON a.[SeqID_Cleaned] + 5 = f.[SeqID_Cleaned]
)

UPDATE CTE SET [HeadingPost5] = (CASE When AlphaPost1 is null or StopPost1 = 1 THEN null Else
   (CASE When AlphaPost2 is null or StopPost2 = 1 THEN AlphaPost1 ELSE
   (CASE When AlphaPost3 is null or StopPost3 = 1 THEN (AlphaPost1 + AlphaPost2) / 2 ELSE
   (CASE When AlphaPost4 is null or StopPost4 = 1 THEN (AlphaPost1 + AlphaPost2 + AlphaPost3) / 3 ELSE
   (CASE When AlphaPost5 is null or StopPost5 = 1 THEN (AlphaPost1 + AlphaPost2 + AlphaPost3 + AlphaPost4) / 4 ELSE
   (AlphaPost1 + AlphaPost2 + AlphaPost3 + AlphaPost4 + AlphaPost5) / 5 END) END) END) END)
From CTE

GO

UPDATE [dbo].[Stops]
SET [HeadingChange] = (CASE When SIGN([HeadingPre5] - [HeadingPost5])
  = 0 then 0 Else
    (CASE When SIGN([HeadingPre5] - [HeadingPost5]) = 1 AND ([HeadingPre5] - [HeadingPost5]) <= 180 then
      ([HeadingPre5] - [HeadingPost5])
    Else
      (CASE When SIGN([HeadingPre5] - [HeadingPost5]) = 1 AND ([HeadingPre5] - [HeadingPost5]) > 180 then
        ([HeadingPost5] + 360 - [HeadingPre5])
      Else
        (CASE When SIGN([HeadingPre5] - [HeadingPost5]) = -1 AND ([HeadingPost5] - [HeadingPre5]) <= 180 then
          ([HeadingPost5] - [HeadingPre5])
        Else
          (CASE When SIGN([HeadingPre5] - [HeadingPost5]) = -1 AND ([HeadingPost5] - [HeadingPre5]) > 180 then
            ([HeadingPre5] + 360 - [HeadingPost5])
          END)END)END)END)END)

GO

<Script 5 End>
Script 6: Calculating the Entropy of Stop Events

Purpose: Calculates entropy pertaining to each stop event. Note that the dataset of stop events can be split into multiple datasets (based on region) and calculated in parallel to speed up the final processing time.

<Script 6 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

CREATE TABLE [dbo].[Entropy_Temp1](
    [CID] [int] NULL,
    [Numerator_Term] [decimal](18, 10) NULL
) ON [PRIMARY]
GO

CREATE TABLE [dbo].[Entropy](
    [SeqID] [bigint] NULL,
    [EntropyIndex] [decimal](5, 2) NULL,
    [Total] [int] NULL,
    [CarrierTotal] [int] NULL
) ON [PRIMARY]
GO

Declare @SEQID as bigint, @CID as varchar(25), @latitude as decimal(18,5), @longitude as decimal(18,5), @COMBO as bigint, @dwellhours as datetime, @Total as decimal(18,5), @geolocation as geography, @latMin as decimal(18,5), @latMax as decimal(18,5), @longMin as decimal(18,5), @longMax as decimal(18,5)

declare db_cursor CURSOR FOR
Select
    [SeqID],
    [Latitude],
    [Longitude],
    [ComboCoordinate],
    [CID],
    [DwellHours]
From [dbo].[Stops]

Open db_cursor
Fetch NEXT From db_cursor into @SEQID, @latitude, @longitude, @COMBO, @CID, @dwellhours

WHILE (@@FETCH_STATUS = 0)
BEGIN
    SET @LatMin = @latitude - 0.00225
    SET @LatMax = @latitude + 0.00225
    SET @LongMin = @longitude - 0.003
    SET @LongMax = @longitude + 0.003

    -- Calculation logic here...

END

212
/* Distances estimated for 250 meters based on 1 degree of latitude = 111.32 km and 1 degree of longitude = 81.4 km (at latitude of 43 degrees) */

```
SET @Total = (SELECT COUNT([SeqID]) FROM [dbo].[Stops]
WHERE [Latitude] >= @LatMin and [Latitude] <= @LatMax and
[Longitude] >= @LongMin and [Longitude] <= @LongMax)

;WITH CTE AS
(SELECT
 [SeqID] = @SEQID,
 [Latitude],
 [Longitude],
 [ComboCoordinate],
 [CID],
 [DwellHours],
 ROW_NUMBER() OVER(ORDER BY [SeqID]) AS RowNumber
 From [dbo].[Stops]
 WHERE [Latitude] >= @LatMin and [Latitude] <= @LatMax and
 [Longitude] >= @LongMin and [Longitude] <= @LongMax
)

INSERT INTO [dbo].[Entropy_Temp1]
SELECT [CID], (COUNT([ComboCoordinate]) / @Total) * LOG(COUNT([ComboCoordinate]) / @Total) as Numerator_Term FROM CTE
GROUP BY [CID]

INSERT INTO [dbo].[Entropy]
SELECT [SeqID] = @SEQID, [EntropyIndex] = -1 * SUM([Numerator_Term]), [Total] = CAST(@Total as bigint), [CarrierTotal] = COUNT([CID])
FROM [dbo].[Entropy_Temp1]

DELETE FROM [dbo].[Entropy_Temp1]
PRINT @SeqID

Fetch NEXT From db_cursor into @SEQID, @latitude, @longitude, @COMBO, @CID, @dwellhours
END
```

Close db_cursor
Deallocate db_cursor
DROP TABLE [dbo].[Entropy_Temp1]
GO

UPDATE a
SET a.[TotalTrucks] = b.[Total],
a.[TotalCarriers] = b.[CarrierTotal],
a.[EntropyIndex] = b.[EntropyIndex]
FROM [Test].dbo.Stops a
LEFT OUTER JOIN [Test].dbo.Entropy b
ON a.SeqID = b.SeqID

SET ANSI_NULLS OFF
GO
/* Creation of Extra Indexes */
USE [GPS_Year2_2014] ---User Input (Database)
GO

CREATE INDEX ix_GPS_Year2_2014_Lat_Long
ON [dbo].[Stops11] ([Latitude], [Longitude]);

CREATE INDEX ix_GPS_Year2_2014_Stops_SeqID
ON [dbo].[Stops11] ([SeqID]);

CREATE INDEX ix_GPS_Year2_2014_CID
ON [dbo].[Stops11] ([CID]);

<Script 6 End>
Script 7: Creation of Dataset for Primary Stops

Purpose: Results in a dataset of stops estimated to be primary (i.e. used to transfer goods).

<Script 7 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

SET ANSI_NULLS ON
GO

CREATE TABLE [dbo].[Stops_Primary](
    [SeqID] [numeric](20, 0) NULL,
    [Latitude] [decimal](28, 5) NULL,
    [Longitude] [decimal](28, 5) NULL,
    [CID] [varchar](50) NULL,
    [CPID] [varchar](50) NULL,
    [TimeEntry] [datetime] NULL,
    [DwellHours] [decimal](28, 5) NULL,
    [TourStart] [varchar](50) NULL,
    [EntropyIndex] [decimal](28, 5) NULL,
    [HeadingChange] [decimal](10, 5) NULL,
    [SIC4] [varchar](50) NULL,
    [SIC2] [varchar](50) NULL,
    [CDUID] [varchar](50) NULL,
    [CDNAME] [varchar](50) NULL,
    [PRUID] [varchar](50) NULL,
    [MSAID] [varchar](50) NULL,
    [CTYID] [varchar](50) NULL,
    [STName] [varchar](50) NULL,
    [STATEPROV] [varchar](10) NULL,
    [DIVISIONID] [varchar](50) NULL
) ON [PRIMARY]
GO

INSERT INTO [dbo].[Stops_Primary]
SELECT [SeqID],
    [Latitude],
    [Longitude],
    [CID],
    [CPID],
    [TimeEntry],
    [DwellHours],
    [TourStart],
    [EntropyIndex],
    [HeadingChange],
    NULL
    ,NULL
    ,NULL
    ,NULL
    ,NULL
    ,NULL
    ,NULL
    ,NULL
    ,NULL
    ,NULL

215
FROM [dbo].[Stops]
Where [EntropyIndex] < 2.75

/* Upon Completion of this script, the data from the primary stops
table needs
to be exported to a .CSV file for upload into:
(1) custom GPS program to identify nearby firms
(2) sorting results in SAS to remove further secondary stops
(3) upload to ArcGIS to determine location of each stop
(4) upload final results back into a table in the SQL database
(below)*/

---This script updates the primary stop table with new information from
an uploaded table and adds indexes for performance gains

USE [Temp_OneDay_Test] ---User Input (Database)
GO

Update a
Set [SIC4] = b.[SIC4],
[SIC2] = CASE(LEN(b.[SIC4]))
        When 4 then LEFT(b.[SIC4],2)
        When 3 then LEFT(b.[SIC4],1)
        END,
[STATEPROV] = b.[Abb],
[DIVISIONID] = b.[TempID]
FROM [GPS_Year2_2014].[dbo].[Stops_Primary] a
    INNER JOIN [CBI_GPS_Mar2016].[dbo].[Input_Information] b ---User
    Input (Updated Stops with Location and Industry)
    ON a.SeqID = b.ID
GO

DELETE FROM [dbo].[Stops_Primary]
WHERE [SIC4] IS NULL
GO

CREATE INDEX ix_Stops_Primary_CPID
ON [dbo].[Stops_Primary] ([CPID]);

CREATE INDEX ix_Stops_Primary_DIVISIONID
ON [dbo].[Stops_Primary] ([DIVISIONID]);

CREATE INDEX ix_Stops_Primary_SeqID
ON [dbo].[Stops_Primary] ([SeqID]);

CREATE INDEX ix_Stops_Primary_SeqID_Desc
ON [dbo].[Stops_Primary] ([SeqID] DESC);

CREATE INDEX ix_Stops_Primary_TimeEntry
ON [dbo].[Stops_Primary] ([TimeEntry]);
GO

<Script 7 End>
Script 8: Identifying Inter-Regional Trips

Purpose: Uses the primary stop event dataset (script 7) as end points to determine inter-regional trips.

<Script 8 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

CREATE TABLE [dbo].[OD_Trips](
    [CarrierID] [varchar](25) NULL,
    [CPIID] [varchar](50) NULL,
    [ORIGINSeqID] [bigint] NULL,
    [ORIGINLatitude] [real] NULL,
    [ORIGINLongitude] [real] NULL,
    [ORIGINStopDwell] [real] NULL,
    [ORIGINExitTime] [datetime] NULL,
    [DESTSeqID] [bigint] NULL,
    [DESTExitTime] [datetime] NULL,
    [DESTDwellTime] [real] NULL,
    [ElapsedTripTime] [real] NULL,
    [ORIGINSIC4] [varchar](4) NULL,
    [ORIGINSIC2] [varchar](4) NULL,
    [DESTSIC4] [varchar](4) NULL,
    [DESTSIC2] [varchar](4) NULL,
    [Origin] [varchar](25) NULL,
    [Destination] [varchar](25) NULL,
    [DESTLatitude] [real] NULL,
    [DESTLongitude] [real] NULL,
    [ORIGINLocation] [geography] NULL,
    [DESTLocation] [geography] NULL,
    [TripKM] [decimal](10, 3) NULL
) ON [PRIMARY]
GO

SET ANSI_PADDING OFF
GO

/* The script below identifies all possible trips between origins and destinations. Keep in mind:
1) This script requires the use of a reference table with all possible census divisions
2) The process can be partitioned into multiple, parallel scripts by adjusting the boundaries for @Origin_ROW
3) Typically, we have also partitioned the process into individual months */

USE [GPS_Year2_2014] ---User Input (Database)
GO
DECLARE @ORIGIN_ROW AS INT, @ORIGIN_NAME AS varchar(15), @DEST_ROW AS INT, @DEST_NAME AS VARCHAR(15)

SET @ORIGIN_ROW = 1  --- Adjust this value for lower bound of partition
SET @DEST_ROW = 1

WHILE @ORIGIN_ROW < 50  --- Adjust this value for upper bound of partition, final value should be 2204
BEGIN

SET @ORIGIN_NAME = (SELECT TOP 1 [DIVISIONID] FROM [GPS_Year_One].[dbo].[CD_Lookup]  ---Reference table providing lookup function for census divisions
WHERE [ROW] = @ORIGIN_ROW)

PRINT @ORIGIN_NAME

SET @DEST_ROW = 1
WHILE @DEST_ROW <= 2203
BEGIN

SET @DEST_NAME = (SELECT TOP 1 [DIVISIONID] FROM [GPS_Year_One].[dbo].[CD_Lookup]
WHERE [ROW] = @DEST_ROW)

IF(@ORIGIN_ROW <> @DEST_ROW)
BEGIN
;WITH CTE AS
(SELECT TOP 1000000000 [CID] AS CarrierID
,[CPID]
,[SeqID] AS ORIGINSeqID
,[Latitude] AS ORIGINLatitude
,[Longitude] AS ORIGINLongitude
,[DwellHours] AS ORIGINStopDwell
,[TimeEntry] AS ORIGINExitTime
ORDER BY [SeqID] DESC) as SeqNum

FROM [dbo].[Stops_Primary_Jan] AS ORIGIN
WHERE ([DIVISIONID] = @ORIGIN_NAME))

INSERT INTO [dbo].[OD Trips_Jan]
SELECT [CarrierID]
,[CPID]
,[ORIGINSeqID]
,[ORIGINLatitude]
,[ORIGINLongitude]
,[ORIGINStopDwell]

FROM CTE

END

END

END
FROM CTE
WHERE SeqNum = 1 AND DESTSeqID IS NOT NULL
END
SET @DEST_ROW = @DEST_ROW + 1
END
SET @ORIGIN_ROW = @ORIGIN_ROW + 1
END
GO

USE [Temp_OneDay_Test] ---User Input (Database)
GO

UPDATE a
SET [DESTExitTime] = c.[TimeEntry]
,[DESTDwellTime] = c.[DwellHours]
,[ElapsedTripTime] = (CAST((c.[TimeEntry] - a.[ORIGINExitTime])
AS Decimal(18,10)) * 24) - c.[DwellHours]
,[ORIGINSIC4] = b.SIC4
,[ORIGINSIC2] = b.SIC2
,[DESTSIC4] = c.SIC4
,[DESTSIC2] = c.SIC2
,[Origin] = b.DIVISIONID
,[Destination] = c.DIVISIONID
,[DESTLatitude] = c.Latitude
,[DESTLongitude] = c.Longitude
FROM [OD_Trips] a
LEFT OUTER JOIN [dbo].[Stops_Primary] b
ON a.ORIGINSeqID = b.SeqID
LEFT OUTER JOIN [dbo].[Stops_Primary] c
ON a.DESTSeqID = c.SeqID

UPDATE a
SET [ORIGINLocation] =
(geography::Point([ORIGINLatitude],[ORIGINLongitude], 4326))
,[DESTLocation] =
(geography::Point([DESTLatitude],[DESTLongitude], 4326))
FROM [dbo].[OD_Trips] a
GO
UPDATE a
SET [TripKM] = a.[ORIGINLocation].STDistance (a.[DESTLocation]) / 1000
FROM [dbo].[OD_Trips] a
GO

SELECT [CarrierID]
,[CPID]
,[ORIGINSeqID]
,[ORIGINLatitude]
,[ORIGINLongitude]
,[ORIGINStopDwell]
,[ORIGINExitTime]
,[Origin]
,[ORIGINIC4]
,[ORIGINIC2]
,[ORIGIN__Detailed_Industry] = ''
,[ORIGIN__General_Industry] = ''
,[DESTSeqID]
,[DESTLatitude]
,[DESTLongitude]
,[DESTDwellTime]
,[DESTExitTime]
,[Destination]
,[DESTSIC4]
,[DESTSIC2]
,[DEST__Detailed_Industry] = ''
,[DEST__General_Industry] = ''
,[Trip Type] = ''
,[Origin_Destination_KM] = [TripKM]
,[ElapsedTripTime]
,[Time_Expected] = ''
,[Time_Max] = ''
,[Valid] = ''
FROM [GPS_Year2_2014].[dbo].[OD_Trips_Feb]

/* Note that some post-processing is still required fill in some blank fields above. This is particularly important for the determination of valid trips based on the elapsed time. */

<Script 8 End>
Script 9: Determining the Canada-U.S. Border Crossing Events

Purpose: This script identifies the trips that cross the Canada-U.S. border at a specific location and provides the crossing time/date of crossing. The script below specifically pertains to the Ambassador Bridge.

<Script 9 Begin>

USE [Temp_OneDay_Test] ---User Input (Database)
GO

--Upper geofence latitude changed from 42.32263 to 42.324 for 2014 data

CREATE TABLE [dbo].[Ambassador_Crossings](
    [CPID] [varchar](50) NULL,
    [Latitude] [real] NULL,
    [Longitude] [real] NULL,
    [TimeEntry] [datetime] NULL,
    [SeqID] [bigint] NULL,
    [SeqID_Cleaned] [bigint] NULL,
    [CrossingNS] [varchar](1) NULL,
    [CrossingTime] [datetime] NULL
)

GO

CREATE TABLE [dbo].[Ambassador_Geofence](
    [CID] [varchar](25) NULL,
    [PID] [varchar](25) NULL,
    [CPID] [varchar](50) NULL,
    [Latitude] [real] NULL,
    [Longitude] [real] NULL,
    [TimeEntry] [datetime] NULL,
    [SeqPID] [bigint] NULL,
    [SeqID] [bigint] NULL,
    [ElapsedTime] [datetime] NULL,
    [DwellTime] [datetime] NULL,
    [Alpha] [decimal](18, 5) NULL,
    [Stop] [int] NULL,
    [SeqID_Cleaned] [bigint] NULL,
    [CrossingNS] [varchar](1) NULL
)

GO

INSERT INTO [dbo].[Ambassador_Geofence]
SELECT
    [CID],
    [PID],
    [CPID],
    [Latitude],
    [Longitude],
    [TimeEntry]

GO
FROM [dbo].[ProcessedData_Oct1_2013] ---User Input (Processed GPS Table)
WHERE [Latitude] >= 42.317142 and [Latitude] <= 42.3244 and
[Longitude] >= -83.081364 and [Longitude] <= -83.076994
or [Latitude] >= 42.308599 and [Latitude] <= 42.317142 and
[Longitude] >= -83.078141 and [Longitude] <= -83.07168
or [Latitude] >= 42.302907 and [Latitude] <= 42.308599 and
[Longitude] >= -83.072792 and [Longitude] <= -83.067267
or [Latitude] >= 42.299501 and [Latitude] <= 42.302907 and
[Longitude] >= -83.068525 and [Longitude] <= -83.062791
GO

Declare @CPID as varchar(50), @latitude as real, @longitude as real,
@TimeEntry as datetime, @SeqID as bigint, @ElapsedTime as datetime,
@DwellTime as Datetime, @Stop as int, @SeqID_Cleaned as bigint, @Start_NS as varchar(1),
@End_NS as varchar(1), @CrossingNS as varchar(1), @leadID as bigint, @lagID as bigint,
@StartLatitude as real, @StartLongitude as real, @EndLatitude as real, @EndLongitude as real,
@StartTime as datetime, @EndTime as datetime,
@N_Entry_Lat as real, @N_Entry_Long as real, @S_Entry_Lat as real,
@S_Entry_Long as real, @Entry_Time as datetime, @Exit_Time as datetime

/* Set coordinates for entry and exit points to Ambassador Geofence (to
determine distances and time interpolations) */
SET @N_Entry_Lat = 42.31989
SET @N_Entry_Long = -83.081364
SET @S_Entry_Lat = 42.299501
SET @S_Entry_Long = -83.065045

PRINT 1
/* Declare cursor to move through points within Ambassador Geofence */
DECLARE db_cursor CURSOR FOR SELECT
  a.[CPID], a.[Latitude], a.[Longitude], a.[TimeEntry], a.[SeqID],
  a.[ElapsedTime], a.[DwellTime], a.[Stop], a.[SeqID_Cleaned],
  a.[CrossingNS], b.[SeqID_Cleaned] as LeadID
, c.[SeqID_Cleaned] as LagID
FROM [dbo].[Ambassador_Geofence] a
LEFT OUTER JOIN [dbo].[Ambassador_Geofence] b
ON a.[SeqID_Cleaned] = b.[SeqID_Cleaned]-1
LEFT OUTER JOIN [dbo].[Ambassador_Geofence] c
ON a.[SeqID_Cleaned] = c.[SeqID_Cleaned]+1
ORDER BY [SeqID_Cleaned]

PRINT 2

Open db_cursor
Fetch NEXT From db_cursor into @CPID, @latitude, @longitude,
@TimeEntry, @SeqID, @elapsedtime, @dwelltime, @Stop, @SeqID_Cleaned,
@CrossingNS, @leadID, @lagID

WHILE (@@FETCH_STATUS = 0)
BEGIN
PRINT 3

/*
Four basic conditions exist for each point inside the Ambassador geofence:
1) No Lead ID or Lag ID - Only one point exists inside the geofence
   Therefore both the entry and exit times are calculated here from the closest two points outside the geofence
   The crossing time is inserted into the result table before fetching the next point

2) No Lag ID - This is the first point inside the geofence, but there are subsequent points within the geofence
   Only the entry time is calculated based on the last point before entering the geofence
   No crossing time is inserted into the result table (the exit time has not been calculated yet) before fetching the next point

3) Lead and Lag points exist - There are previous points and subsequent points inside the geofence
   The entry time is carried over from condition 2
   The exit time has not been calculated yet
   No action is taken and the next point is fetched

4) No Lead ID - This is the last point inside the geofence, but there were previous points within the geofence
   The entry time is carried over from condition 2
   Only the exit time is calculated based on the next point after leaving the geofence
   The crossing time is inserted into the result table before fetching the next point
*/

/** CONDITION 1 **/
IF(@LeadID is NULL AND @LagID is NULL)
BEGIN
Print 4
/* Set the time and location for the last point before entering the geofence */
SET @StartTime = (SELECT [TimeEntry] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned - 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @StartLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned - 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @StartLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned - 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

/* Set the time and location for the first point inside the geofence */
SET @FirstTime = (SELECT [TimeEntry] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = @SeqID_Cleaned AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @FirstLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = @SeqID_Cleaned AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @FirstLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = @SeqID_Cleaned AND [CPID] = @CPID) ---User Input (Processed GPS Table)

/* Determine if the truck is starting in the North (US) or South (Canada) */
SET @Start_NS = NULL

IF((@StartLatitude BETWEEN 42.24 AND 42.375 AND @StartLongitude BETWEEN -83.17 AND -83.113) OR (@StartLatitude BETWEEN 42.291 AND 42.375 AND @StartLongitude BETWEEN -83.113 AND -83.095) OR (@StartLatitude BETWEEN 42.301 AND 42.375 AND @StartLongitude BETWEEN -83.095 AND -83.085) OR (@StartLatitude BETWEEN 42.312 AND 42.375 AND @StartLongitude BETWEEN -83.085 AND -83.069) OR (@StartLatitude BETWEEN 42.322 AND 42.375 AND @StartLongitude BETWEEN -83.069 AND -83.039) OR (@StartLatitude BETWEEN 42.33 AND 42.375 AND @StartLongitude BETWEEN -83.039 AND -82.99))
BEGIN
  SET @Start_NS = 'N'
  SET @Distance_Outside_1 = SQRT(ABS(@StartLatitude-@N_Entry_Lat) + ABS(@StartLongitude-@N_Entry_Long))
  SET @Distance_Inside_1 = SQRT(ABS(@FirstLatitude-@N_Entry_Lat) + ABS(@FirstLongitude-@N_Entry_Long))
END

IF((@StartLatitude BETWEEN 42.24 AND 42.291 AND @StartLongitude BETWEEN -83.113 AND -83.095) OR (@StartLatitude BETWEEN 42.24 AND 42.301 AND @StartLongitude BETWEEN -83.095 AND -83.085) OR (@StartLatitude BETWEEN 42.24 AND 42.312 AND @StartLongitude BETWEEN -83.085 AND -83.069) OR (@StartLatitude BETWEEN 42.24 AND 42.322 AND @StartLongitude BETWEEN -83.069 AND -83.039) OR (@StartLatitude BETWEEN 42.24 AND 42.33 AND @StartLongitude BETWEEN -83.039 AND -82.99))
BEGIN
SET @Start_NS = 'S'
SET @Distance_Outside_1 = SQRT(ABS(@StartLatitude - @S_Entry_Lat) + ABS(@StartLongitude - @S_Entry_Long))
SET @Distance_Inside_1 = SQRT(ABS(@FirstLatitude - @S_Entry_Lat) + ABS(@FirstLongitude - @S_Entry_Long))
END

/*Remove the start designation if the point before entering the geofence is in an erroneous zone (such as the water) */
IF((@StartLatitude BETWEEN 42.30291 AND 42.31714 AND @StartLongitude BETWEEN -83.0792 AND -83.06383) OR (@StartLatitude BETWEEN 42.29881 AND 42.31414 AND @StartLongitude BETWEEN -83.085 AND -83.0792) OR (@StartLatitude BETWEEN 42.31414 AND 42.31548 AND @StartLongitude BETWEEN -83.08187 AND -83.0792) OR (@StartLatitude BETWEEN 42.31548 AND 42.31649 AND @StartLongitude BETWEEN -83.08044 AND -83.0792) OR (@StartLatitude BETWEEN 42.31649 AND 42.31697 AND @StartLongitude BETWEEN -83.07956 AND -83.0792) OR (@StartLatitude BETWEEN 42.29472 AND 42.302907 AND @StartLongitude BETWEEN -83.07916 AND -83.069) OR (@StartLatitude BETWEEN 42.317142 AND 42.32263 AND @StartLongitude BETWEEN -83.07569 AND -83.069))
BEGIN
SET @Start_NS = NULL
END

/* Determine the entry time into the geofence (using a linear interpolation) */
SET @Entry_Time = @FirstTime - Cast(@Distance_Inside_1 / (SQRT(ABS(@FirstLatitude - @StartLatitude) + ABS(@FirstLongitude - @StartLongitude))/ CAST(@FirstTime - @StartTime as decimal(18,6))) as datetime)

/* If the last point before entering the geofence is closer to the first point in the geofence then the entry point OR if the First and Start times are the same -- the entry time is set to the point before entering the geofence. */
IF((@Entry_Time < @StartTime) OR @FirstTime = @StartTime)
BEGIN
SET @Entry_Time = @StartTime
END

/* Determine if the truck is ending in the North (US) or South (Canada) */
SET @Endtime = (SELECT [TimeEntry] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned + 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @EndLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned + 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @EndLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned + 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @LastTime = (SELECT [TimeEntry] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @LastLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @LastLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @End_NS = NULL

IF((@EndLatitude BETWEEN 42.24 AND 42.375 AND @EndLongitude BETWEEN -83.17 AND -83.113) OR (@EndLatitude BETWEEN 42.291 AND 42.375 AND @EndLongitude BETWEEN -83.113 AND -83.095) OR (@EndLatitude BETWEEN 42.301 AND 42.375 AND @EndLongitude BETWEEN -83.095 AND -83.085) OR (@EndLatitude BETWEEN 42.312 AND 42.375 AND @EndLongitude BETWEEN -83.085 AND -83.069) OR (@EndLatitude BETWEEN 42.322 AND 42.375 AND @EndLongitude BETWEEN -83.069 AND -83.039) OR (@EndLatitude BETWEEN 42.33 AND 42.375 AND @EndLongitude BETWEEN -83.039 AND -82.99))

BEGIN
SET @End_NS = 'N'
SET @Distance_Outside_2 = SQRT(ABS(@EndLatitude - @N_Entry_Lat) + ABS(@EndLongitude - @N_Entry_Long))
SET @Distance_Inside_2 = SQRT(ABS(@LastLatitude - @N_Entry_Lat) + ABS(@LastLongitude - @N_Entry_Long))
END

IF((@EndLatitude BETWEEN 42.24 AND 42.291 AND @EndLongitude BETWEEN -83.113 AND -83.095) OR (@EndLatitude BETWEEN 42.24 AND 42.301 AND @EndLongitude BETWEEN -83.095 AND -83.085) OR (@EndLatitude BETWEEN 42.24 AND 42.312 AND @EndLongitude BETWEEN -83.085 AND -83.069) OR (@EndLatitude BETWEEN 42.24 AND 42.322 AND @EndLongitude BETWEEN -83.069 AND -83.039) OR (@EndLatitude BETWEEN 42.24 AND 42.33 AND @EndLongitude BETWEEN -83.039 AND -82.99))

BEGIN
SET @End_NS = 'S'
SET @Distance_Outside_2 = SQRT(ABS(@EndLatitude - @S_Entry_Lat) + ABS(@EndLongitude - @S_Entry_Long))
SET @Distance_Inside_2 = SQRT(ABS(@LastLatitude - @S_Entry_Lat) + ABS(@LastLongitude - @S_Entry_Long))
END

IF((@EndLatitude BETWEEN 42.30291 AND 42.31714 AND @EndLongitude BETWEEN -83.0792 AND -83.06383) OR
(@EndLatitude BETWEEN 42.29881 AND 42.31414 AND
@EndLongitude BETWEEN -83.085 AND -83.0792) OR
(@EndLatitude BETWEEN 42.31414 AND 42.31548 AND
@EndLongitude BETWEEN -83.0844 AND -83.0792) OR
(@EndLatitude BETWEEN 42.31548 AND 42.31649 AND
@EndLongitude BETWEEN -83.08044 AND -83.0792) OR
(@EndLatitude BETWEEN 42.31649 AND 42.31697 AND
@EndLongitude BETWEEN -83.07956 AND -83.0792) OR
(@EndLatitude BETWEEN 42.29472 AND 42.302907 AND
@EndLongitude BETWEEN -83.07916 AND -83.069) OR
(@EndLatitude BETWEEN 42.317142 AND 42.32263 AND
@EndLongitude BETWEEN -83.07569 AND -83.069)
BEGIN
    SET @End_NS = NULL
END

/* Determine the exit time from the geofence based on a
linear interpolation */
SET @Exit_Time = @LastTime + CAST(@Distance_Inside_2 /
(SQRT(ABS(@LastLatitude-@EndLatitude)+ABS(@EndLongitude-
@LastLongitude)) / CAST(@EndTime - @LastTime as decimal(18,6))) as
datetime)

IF((@Exit_Time > @EndTime) OR (@LastTime = @EndTime))
BEGIN
    SET @Exit_Time = @EndTime
END

Print @End_NS
PRINT @Start_NS
/* Insert the crossing time into the results table if the
truck started and ended in a different country */
IF((@End_NS = 'N' AND @Start_NS = 'S') OR (@End_NS = 'S'
AND @Start_NS = 'N'))
BEGIN
    INSERT INTO [dbo].[Ambassador_Crossings]
    (CPID,Latitude,Longitude,TimeEntry,SeqID,SeqID_Cleaned,CrossingNS
,CrossingTime)
    VALUES(@CPID, @latitude, @longitude, @Entry_Time,
@SeqID, @SeqID_Cleaned, @End_NS, @Exit_Time - @Entry_Time)
END

Fetch NEXT From db_cursor into @CPID, @latitude,
@longitude, @TimeEntry, @SeqID, @elapsedtime, @dwelltime, @Stop,
@SeqID_Cleaned, @CrossingNS, @leadID, @lagID
END
ELSE
/** CONDITION 2 **/
IF(@LagID is NULL)
Begin
Print 5
Set @StartTime = (SELECT [TimeEntry] FROM
[dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned - 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @StartLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned - 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @StartLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned - 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @FirstTime = (SELECT [TimeEntry] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = @SeqID_Cleaned AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @FirstLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = @SeqID_Cleaned AND [CPID] = @CPID) ---User Input (Processed GPS Table)
SET @FirstLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = @SeqID_Cleaned AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @Start_NS = NULL

IF((@StartLatitude BETWEEN 42.24 AND 42.375 AND @StartLongitude BETWEEN -83.17 AND -83.113) OR (@StartLatitude BETWEEN 42.291 AND 42.375 AND @StartLongitude BETWEEN -83.113 AND -83.095) OR (@StartLatitude BETWEEN 42.301 AND 42.375 AND @StartLongitude BETWEEN -83.095 AND -83.085) OR (@StartLatitude BETWEEN 42.312 AND 42.375 AND @StartLongitude BETWEEN -83.085 AND -83.069) OR (@StartLatitude BETWEEN 42.322 AND 42.375 AND @StartLongitude BETWEEN -83.069 AND -83.039) OR (@StartLatitude BETWEEN 42.33 AND 42.375 AND @StartLongitude BETWEEN -83.039 AND -82.99))
BEGIN
SET @Start_NS = 'N'
SET @Distance_Outside_1 = SQRT(ABS(@StartLatitude-@N_Entry_Lat) + ABS(@StartLongitude-@N_Entry_Long))
SET @Distance_Inside_1 = SQRT(ABS(@FirstLatitude-@N_Entry_Lat) + ABS(@FirstLongitude-@N_Entry_Long))
END

IF((@StartLatitude BETWEEN 42.24 AND 42.291 AND @StartLongitude BETWEEN -83.113 AND -83.095) OR (@StartLatitude BETWEEN 42.24 AND 42.301 AND @StartLongitude BETWEEN -83.095 AND -83.085) OR (@StartLatitude BETWEEN 42.24 AND 42.312 AND @StartLongitude BETWEEN -83.085 AND -83.069) OR (@StartLatitude BETWEEN 42.24 AND 42.322 AND @StartLongitude BETWEEN -83.069 AND -83.039) OR (@StartLatitude BETWEEN 42.24 AND 42.33 AND @StartLongitude BETWEEN -83.039 AND -82.99))
BEGIN
SET @Start_NS = 'S'
SET @Distance_Outside_1 = SQRT(ABS(@StartLatitude-@S_Entry_Lat) + ABS(@StartLongitude-@S_Entry_Long))
SET @Distance_Inside_1 = SQRT(ABS(@FirstLatitude-@S_Entry_Lat) + ABS(@FirstLongitude-@S_Entry_Long))
END
IF((@StartLatitude BETWEEN 42.30291 AND 42.31714 AND
@StartLongitude BETWEEN -83.0792 AND -83.06383) OR
(@StartLatitude BETWEEN 42.29881 AND 42.31414 AND
@StartLongitude BETWEEN -83.085 AND -83.06383) OR
(@StartLatitude BETWEEN 42.31414 AND 42.31548 AND
@StartLongitude BETWEEN -83.08187 AND -83.06383) OR
(@StartLatitude BETWEEN 42.31548 AND 42.31649 AND
@StartLongitude BETWEEN -83.08044 AND -83.06383) OR
(@StartLatitude BETWEEN 42.31649 AND 42.31697 AND
@StartLongitude BETWEEN -83.07956 AND -83.06383) OR
(@StartLatitude BETWEEN 42.29472 AND 42.302907 AND
@StartLongitude BETWEEN -83.07916 AND -83.069) OR
(@StartLatitude BETWEEN 42.317142 AND 42.32263 AND
@StartLongitude BETWEEN -83.07569 AND -83.069))
BEGIN
  SET @Start_NS = NULL
END

SET @Entry_Time = @FirstTime - Cast(@Distance_Inside_1 /
(SQRT(ABS(@FirstLatitude-@StartLatitude)+ABS(@FirstLongitude-
@StartLongitude))/ CAST(@FirstTime - @StartTime as decimal(18,6))) as
datetime)

IF((@Entry_Time < @StartTime) OR @FirstTime = @StartTime)
BEGIN
  SET @Entry_Time = @StartTime
END

Fetch NEXT From db_cursor into @CPID, @latitude,
@longitude, @TimeEntry, @SeqID, @elapsedtime, @dwelltime, @Stop,
(SeqID_Cleaned, @CrossingNS, @leadID, @lagID
END

ELSE
/** CONDITION 3 **/
  IF(@LeadID > 0 and @LagID > 0)
  BEGIN
    Print 6
    Fetch NEXT From db_cursor into @CPID, @latitude,
    @longitude, @TimeEntry, @SeqID, @elapsedtime, @dwelltime, @Stop,
    @SeqID_Cleaned, @CrossingNS, @leadID, @lagID
  END

ELSE
/** CONDITION 4 **/
  IF(@LeadID is NULL)
  BEGIN
    Print 7
    SET @Endtime = (SELECT [TimeEntry] FROM
    [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned
    + 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
    SET @EndLatitude = (SELECT [Latitude] FROM
    [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned
    + 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)
    SET @EndLongitude = (SELECT [Longitude] FROM
    [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned
    + 1) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

    ELSE
*/
SET @LastTime = (SELECT [TimeEntry] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @LastLatitude = (SELECT [Latitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @LastLongitude = (SELECT [Longitude] FROM [dbo].[ProcessedData_Oct1_2013] WHERE [SeqID_Cleaned] = (@SeqID_Cleaned) AND [CPID] = @CPID) ---User Input (Processed GPS Table)

SET @End_NS = NULL

IF(@EndLatitude BETWEEN 42.24 AND 42.375 AND @EndLongitude BETWEEN -83.17 AND -83.113 OR @EndLatitude BETWEEN 42.291 AND 42.375 AND @EndLongitude BETWEEN -83.113 AND -83.095 OR @EndLatitude BETWEEN 42.301 AND 42.375 AND @EndLongitude BETWEEN -83.095 AND -83.085 OR @EndLatitude BETWEEN 42.312 AND 42.375 AND @EndLongitude BETWEEN -83.085 AND -83.069 OR @EndLatitude BETWEEN 42.322 AND 42.375 AND @EndLongitude BETWEEN -83.069 AND -83.039 OR @EndLatitude BETWEEN 42.33 AND 42.375 AND @EndLongitude BETWEEN -83.039 AND -82.99)
BEGIN
    SET @End_NS = 'N'
    SET @Distance_Outside_2 = SQRT(ABS(@EndLatitude-@N_Entry_Lat) + ABS(@EndLongitude-@N_Entry_Long))
    SET @Distance_Inside_2 = SQRT(ABS(@LastLatitude-@N_Entry_Lat) + ABS(@LastLongitude-@N_Entry_Long))
END

IF(@EndLatitude BETWEEN 42.24 AND 42.291 AND @EndLongitude BETWEEN -83.113 AND -83.095 OR @EndLatitude BETWEEN 42.24 AND 42.301 AND @EndLongitude BETWEEN -83.095 AND -83.085 OR @EndLatitude BETWEEN 42.24 AND 42.312 AND @EndLongitude BETWEEN -83.085 AND -83.069 OR @EndLatitude BETWEEN 42.24 AND 42.322 AND @EndLongitude BETWEEN -83.069 AND -83.039 OR @EndLatitude BETWEEN 42.24 AND 42.33 AND @EndLongitude BETWEEN -83.039 AND -82.99)
BEGIN
    SET @End_NS = 'S'
    SET @Distance_Outside_2 = SQRT(ABS(@EndLatitude-@S_Entry_Lat) + ABS(@EndLongitude-@S_Entry_Long))
    SET @Distance_Inside_2 = SQRT(ABS(@LastLatitude-@S_Entry_Lat) + ABS(@LastLongitude-@S_Entry_Long))
END

IF((@EndLatitude BETWEEN 42.30291 AND 42.31714 AND @EndLongitude BETWEEN -83.0792 AND -83.06383) OR
BEGIN
  SET @End_NS = NULL
END

SET @Exit_Time = @LastTime + CAST(@Distance_Inside_2 /
  (SQRT(ABS(@LastLatitude-@EndLatitude)+ABS(@EndLongitude-
  @LastLongitude)) / CAST(@EndTime - @LastTime as decimal(18,6))) as
datetime)

IF((@Exit_Time > @EndTime) OR (@LastTime = @EndTime))
BEGIN
  SET @Exit_Time = @EndTime
END

IF((@End_NS = 'N' AND @Start_NS = 'S') OR (@End_NS = 'S'
  AND @Start_NS = 'N'))
BEGIN
  INSERT INTO [dbo].[Ambassador_Crossings]
  (CPID, Latitude, Longitude, TimeEntry, SeqID, SeqID_Cleaned, CrossingNS,
  CrossingTime)
  VALUES (@CPID, @latitude, @longitude, @Entry_Time, @SeqID,
  @SeqID_Cleaned, @End_NS, @Exit_Time - @Entry_Time)
END

Fetch NEXT From db_cursor into @CPID, @latitude, @longitude,
@TimeEntry, @SeqID, @elapsedtime, @dwelltime, @Stop, @SeqID_Cleaned,
@CrossingNS, @leadID, @lagID
END

Close db_cursor
Deallocate db_cursor
GO

<Script 9 End>
VITA AUCTORIS

NAME: Kevin Gingerich
PLACE OF BIRTH: Kitchener, ON
YEAR OF BIRTH: 1987
EDUCATION: Waterloo-Oxford High School, Baden, ON, 2005
University of Windsor, B.A.Sc., Windsor, ON, 2009
University of Windsor, M.A.Sc., Windsor, ON, 2013
University of Windsor, PhD, Windsor, ON, 2017