Modelling Intercity Travel Demand in the Quebec City – Windsor Corridor: A Study on the Effects of Local Accessibility Utilizing a Geographically Disaggregated Web-Based Approach

by

Billy Wong

A thesis submitted in conformity with the requirements for the degree of M.A.Sc
Department of Civil Engineering
University of Toronto

© Copyright by Billy Wong 2013
Modelling Intercity Travel Demand in the Quebec City – Windsor Corridor: A Study on the Effects of Local Accessibility Utilizing a Geographically Disaggregated Web-Based Approach

Billy Wong

M.A.Sc

Department of Civil Engineering
University of Toronto

2013

Abstract

This thesis presents a novel approach to evaluating intercity travel demand given the introduction of a new travel mode. A joint RP-SP survey is designed to collect both experienced and hypothetical travel decisions. The stated preference aspect of the survey evaluates the level-of-service attributes that influence intercity mode choice. Emphasis is placed on the effects of local accessibility by adapting a geographical disaggregation at the Forward Sortation Area (FSA). Survey distribution collected data primarily through the use of online social networks such as Facebook and Reddit in addition to an online marketing research survey panel. Intercity mode choice was modeled with Random Utility Maximizing (RUM) discrete choice models and is represented by Multinomial Logit (MNL) and Nested Logit (NL) choice structures. Key finding is that access to transit stations significantly influences intercity mode choices.
Acknowledgments

Many thanks go towards Khandker Nurul Habib for his advice and support in the past two years, your enthusiasm and knowledge about transportation demand drives me to constantly improve my own knowledge and understanding. Also a many thanks go out to all the other graduate students in the transportation group for all the smiles and cries. I would not know how I could have gone this far without you guys.

Outside of campus, I would like to thank Michelle and my mom for their support; even though you’re far away you have always been a phone call away.
# Table of Contents

Acknowledgments ........................................................................................................ iii

Table of Contents ........................................................................................................ iv

List of Tables ............................................................................................................... ix

List of Figures ............................................................................................................. xi

List of Appendices ...................................................................................................... xiv

Chapter 1 Introduction ............................................................................................... 1

1  Introduction ............................................................................................................. 1

1.1  Objective ............................................................................................................. 4

1.2  Methods .............................................................................................................. 4

1.3  Thesis Outline .................................................................................................. 6

Chapter 2 Literature Review ..................................................................................... 7

2  Overview ............................................................................................................... 7

2.1  Quebec City – Windsor Corridor .................................................................. 7

2.2  Intercity Travel Demand .............................................................................. 9

2.2.1  Existing Canadian Travel Demand Models .......................................... 10

2.2.2  Existing International Travel Demand Models .................................... 13

2.3  High Speed Rail .............................................................................................. 17

2.3.1  Ecotrain Feasibility Study ................................................................. 19

2.3.2  RAND Corporation Feasibility Study .................................................. 25

2.4  Web-based Survey ......................................................................................... 29

2.5  Web Recruitment ............................................................................................ 31

2.6  Literature Review Summary .......................................................................... 38

Chapter 3 Survey Design ......................................................................................... 39
3 Introduction ........................................................................................................................................... 39

3.1 Intercity Revealed Preference .................................................................................................................. 39

3.2 Stated Preference ........................................................................................................................................ 40

3.2.1 Background ........................................................................................................................................... 40

3.2.2 Hypothetical Situation ............................................................................................................................ 40

3.2.3 Procedure .............................................................................................................................................. 41

3.2.4 Alternatives ........................................................................................................................................... 41

3.2.5 Attribute and Levels ............................................................................................................................... 46

3.2.6 Selected Attributes and Levels ............................................................................................................. 47

3.2.7 Preliminary Utility Function .................................................................................................................. 62

3.2.8 Ngene Software .................................................................................................................................... 63

3.3 Local Revealed Preference ......................................................................................................................... 65

3.4 Demographics .......................................................................................................................................... 66

Chapter 4 Survey Build Setup .......................................................................................................................... 68

4 Introduction ................................................................................................................................................. 68

4.1 Forward Sortation Area .............................................................................................................................. 68

4.1.1 Problem ............................................................................................................................................... 69

4.1.2 Data Source ......................................................................................................................................... 70

4.1.3 Postal Code versus FSA ....................................................................................................................... 71

4.1.4 Problems and Limitations .................................................................................................................... 73

4.2 Google Maps ............................................................................................................................................ 74

4.2.1 Adapting FSA Coordinates .................................................................................................................. 74

4.2.2 Query Travel Data ................................................................................................................................ 77

4.2.3 Benefits ............................................................................................................................................... 78

4.2.4 Limitations ........................................................................................................................................... 79

4.3 Excel ....................................................................................................................................................... 79
4.3.1  Rational....................................................................................................................... 80
4.3.2  Stated Preference Setup ........................................................................................... 81
4.3.3  Data Analysis Setup ................................................................................................. 98
4.4  AppleScript .................................................................................................................. 102

Chapter 5 Web Survey Design ...................................................................................... 104
5  Introduction ..................................................................................................................... 104
  5.1  Alternatives .................................................................................................................. 104
    5.1.1  Coding from scratch ............................................................................................... 105
    5.1.2  Web-based Service ................................................................................................. 107
  5.2  Survey Instrument Selection ...................................................................................... 113
    5.2.1  Benefits .................................................................................................................. 113
    5.2.2  Limitations ............................................................................................................. 116
  5.3  Adaptation of External Coding .................................................................................. 117
  5.4  Survey Flow ................................................................................................................ 122

Chapter 6 Sample Selection Methodology .................................................................... 124
6  Introduction ..................................................................................................................... 124
  6.1  Target Population ........................................................................................................ 124
    6.1.1  Population Count .................................................................................................. 124
    6.1.2  Prior Intercity Trip ................................................................................................. 126
  6.2  Sample Frame ............................................................................................................. 127
  6.3  Sample Recruitment .................................................................................................... 128
    6.3.1  Online Social Media ............................................................................................. 129
    6.3.2  Survey Panel ......................................................................................................... 130
    6.3.3  Phone Lists ........................................................................................................... 130
    6.3.4  Street Intercept .................................................................................................... 131
    6.3.5  General Advertising ............................................................................................. 131
# Table of Contents

Chapter 7 Survey Observations ................................................................. 133

7 Introduction ......................................................................................... 133
  7.1 Pilot Survey .................................................................................... 133
  7.2 Final Survey .................................................................................. 134
    7.2.1 Collector Sources .................................................................... 135
    7.2.2 SurveyMonkey Panel Data Trends ........................................... 148
    7.2.3 Non-Panel Data Trends ............................................................. 155
    7.2.4 Validation of Collected Data ..................................................... 164

Chapter 8 Data Analysis ........................................................................... 167

8 Introduction ......................................................................................... 167
  8.1 Modelling Framework ..................................................................... 167
    8.1.1 Multinomial Logit .................................................................... 168
    8.1.2 Nested Logit ............................................................................ 169
  8.2 Pilot Survey .................................................................................... 169
  8.3 Final Survey ................................................................................... 171
    8.3.1 Multinomial Logit .................................................................... 172
    8.3.2 Nested Logit – Single Nested .................................................. 174
    8.3.3 Nested Logit – Double Nested ................................................ 176
    8.3.4 Summary ................................................................................. 179
  8.4 Model Validation ............................................................................ 181
  8.5 Value of Travel Time Savings .......................................................... 182
  8.6 Sensitivity Analysis ......................................................................... 184

Chapter 9 Conclusion ................................................................................ 188

9 Introduction ......................................................................................... 188
  9.1 Survey Design ............................................................................... 188
  9.2 Web-based Build ............................................................................ 189
9.3 Data Collection ................................................................. 190
9.4 Demand Modelling .......................................................... 192
9.5 Future Work ................................................................. 193
References ............................................................................. 194
Appendices ............................................................................. 1
List of Tables

Table 3-1 – 2006 Annual Non-Auto Trips in the Quebec City – Windsor Corridor .................... 21
Table 3-2 – 2009 License Plate Survey Results............................................................................. 22
Table 3-3 – Stated Preference Survey Results............................................................................... 22
Table 3-4 – Weight Calculation for Model Estimation ................................................................. 24
Table 3-5 – Fare Elasticity of Business Model for Quebec to Montreal Segment ..................... 24
Table 3-6 – Fare Elasticity of Non-Business Model for Quebec to Montreal Segment ............ 25
Table 3-7 – In Vehicle Travel Time of Business Model for Quebec to Montreal Segment ....... 25
Table 3-8 – In Vehicle Travel Time of Non-Business Model for Quebec to Montreal Segment. 25
Table 3-9 – Breakdown of SP Interviews by Mode and Survey Approach............................... 26
Table 3-10 – Breakdown of SP Interviews by Mode and Survey Approach.............................. 26
Table 3-11 – Summary of Examined Attributes in SP Choice Experiment ............................... 28
Table 4-1 – Baseline values for each transportation mode .......................................................... 42
Table 4-2 – Average Annual Operating Costs per Kilometer - Variable ..................................... 42
Table 4-3 – Travel Times Between Major City Pairs ................................................................. 45
Table 4-4 – Number of Train sets per Direction per Day ............................................................ 46
Table 4-5 – Attributes considered in stated preference survey .................................................. 48
Table 5-1 – Verified Postal Codes From geocoder.ca Dataset .................................................. 71
Table 6-1 – List of Web-Based Survey Services Considered....................................................... 105
Table 7-1 – Summary of 2011 Census Population and Dwelling Count ................................. 126
Table 8-1 – Summary of Data Collection Program ................................................................. 148
Table 9-1 – Results of Biogeme multinomial logit model .............................................................. 170

Table 9-2 – Summary of Model Estimation Outputs from STATA 10 ........................................ 181

Table 9-3 – Validation 1 and 2 .................................................................................................... 182

Table 9-4 – Table of Value of Time Travel Savings .................................................................... 183
List of Figures

Figure 3-1 – Map of the Quebec City – Windsor Corridor.......................................................... 8
Figure 3-2 – Sample SP Scenario in EcoTrain Survey .................................................................... 23
Figure 3-3 - Sample SP Scenario in RAND Study Survey ............................................................. 27
Figure 3-4 – Nesting Structure of RAND Study’s Nesting Logit Model ....................................... 29
Figure 5-1 – Calculated Centroid Location of M5T FSA ............................................................... 73
Figure 5-2 – FSA Boundaries from Canada Post .............................................................................. 73
Figure 5-3 – Screenshot of Spreadsheet Mapper 3 layout ............................................................. 76
Figure 5-4 – Map of FSA Centroid Locations and City Boundaries ............................................. 77
Figure 5-4 – Flow of Data in Microsoft Excel .................................................................................. 82
Figure 5-5 – Comparison between NgeneRAW and Ngene .......................................................... 87
Figure 5-6 – Comparison of Stated Preference Tables ................................................................. 94
Figure 5-7 - Sample Stated Preference Scenario in Proposed Survey ......................................... 97
Figure 6-1 – Hierarchy of Information Stored on Logaday Server .............................................. 119
Figure 8-1 – Frequency of Provincial Travel (Panel) ...................................................................... 149
Figure 8-2 – Destination of Last Intercity Trip in the QWC (Panel) ............................................. 150
Figure 8-3 – Purpose of Last Intercity Trip in the QWC (Panel) .................................................... 150
Figure 8-4 – Travel Mode of Last Intercity Trip in the QWC (Panel) ........................................... 151
Figure 8-5 – Purpose of Hypothetical Stated Preference Trip (Panel) ........................................... 151
Figure 8-6 – Purpose of Daily Travel (Panel) .................................................................................. 152
Figure 8-7 – Age Distribution of Respondents (Panel) ................................................................. 153
Figure 8-8 – Gender of Survey Respondents (Panel) ................................................................. 153
Figure 8-9 – Status of Survey Respondents (Panel) ................................................................. 153
Figure 8-10 – Household Size (Panel) ..................................................................................... 154
Figure 8-11 – Household Auto Ownership (Panel) .................................................................. 154
Figure 8-12 – Household Income (Panel) .................................................................................. 155
Figure 8-13 – Frequency of Provincial Travel (Non-Panel) ....................................................... 156
Figure 8-14 – Destination of Last Intercity Trip in the QWC (Non-Panel) ............................... 156
Figure 8-15 – Purpose of Last Intercity Trip in the QWC (Non-Panel) ..................................... 157
Figure 8-16 – Travel Mode of Last Intercity Trip in the QWC (Non-Panel) .............................. 158
Figure 8-17 – Purpose of Hypothetical Stated Preference Trip (Non-Panel) ............................ 158
Figure 8-18 – Purpose of Daily Travel (Non-Panel) ................................................................ 159
Figure 8-19 – Age Distribution of Respondents (Non-Panel) .................................................. 160
Figure 8-20 – Gender of Survey Respondents (Non-Panel) ...................................................... 161
Figure 8-21 – Status of Survey Respondents (Non-Panel) ....................................................... 161
Figure 8-22 – Household Size (Non-Panel) ............................................................................ 162
Figure 8-23 – Household Auto Ownership (Non-Panel) .......................................................... 163
Figure 8-24 – Household Income (Non-Panel) ....................................................................... 163
Figure 8-25 – Distribution of population, dwellings, and survey observations ....................... 165
Figure 9-1 – Nested Logit Structure ....................................................................................... 175
Figure 9-2 – Double-Nested Logit Structure .......................................................................... 177
Figure 9-3 – Sensitivity W.R.T. Changing HSR Costs.......................................................... 184

Figure 9-4 – Sensitivity W.R.T. Changing Costs for All Alternatives .................................. 185

Figure 9-5 – Sensitivity W.R.T. Changing HSR Travel Time............................................... 186

Figure 9-6 – Sensitivity W.R.T. Changing Travel Time for All Alternatives ......................... 187
List of Appendices

Appendix A - EcoTrain Survey
Appendix B - EcoTrain Nesting Structure
Appendix C - SurveyMonkey Questions
Appendix D - Ngene Input Code
Appendix E - Postal Code Map
Appendix F - GeoMidpoint Calculation Methods
Appendix G - Average Longitude & Latitude Calculation Steps
Appendix H - Excel Worksheet Examples
Appendix I - Preliminary SP Procedures
Appendix J - Index Match Function Procedure
Appendix K - Travel Time Calculation Procedure
Appendix L - HTML Output Code Example
Appendix M - AppleScript Code
Appendix N - Demographics by Collector
Appendix O - Geographic Spread of Respondents
Appendix P - Stata Input Code
Appendix Q - Stata Outputs
Chapter 1
Introduction

The continued economic growth and development of urban areas within Canada have created corridors of population, economic activity, and movement of individuals. Movement of individuals within these urban corridors has become the subject of numerous studies from economics, social science, and transportation research groups. Within federal, provincial, and municipal transportation development sectors, considerable efforts are dedicated to assessing the different modes of passenger and freight movement within these corridors (IBI Group, 2002).

With increased passenger volumes travelling within these corridors, alternate transportation modes have been proposed and researched in the past without implementation. One such project is a high speed rail system (HSR), which has been studied since the early 90’s (Langan, 2011). Increasing oil price has been one of the motivations for continued research and development; however, the necessity to accommodate for multiple political agencies has been one of the barriers towards approval of these intercity transportation projects (Miller, 2004). The primary proposed transportation alternative assessed in this project is a high speed rail system for passenger and freight transport spanning between Quebec-City to Windsor.

In the most recent feasibility study regarding a high speed rail system in the Quebec City – Windsor corridor (QWC), a consortium of consulting firms under the name EcoTrain was asked by Transport Canada, Ministry of Transportation of Ontario, and Ministere des Transports du Quebec to comprehensively assess the application of a high speed rail line for operating in 2025 (EcoTrain, 2011). A combined revealed preference (RP) and stated preference (SP) survey was designed and distributed through intercept and license survey collection methods. The conclusion of the EcoTrain report stipulated that a HSR route would attract more than 10 million passengers by 2031, but would only be economically sustainable for key route segments along the QWC (EcoTrain, 2011).

A similar study on high speed rail was completed by the RAND Corporation on the feasibility of a high speed rail line operating in Great Britain. Similar to the EcoTrain feasibility study, this study also utilized a combined RP-SP survey and collected data mainly via intercepting
respondents mid-journey. The resulting model suggested that the value placed on HSR, over traditional rail, differs based on the mode of travel that the respondent was using during the survey. There are often issues with collected datasets as many consulting firms do not publically disclose collected datasets. Compared to the availability of data and models on intracity (local) travel demand, there is currently a gap in the transportation research field regarding intercity travel demand. Current publically available data sets on Canadian intercity travel lack local access and egress information, which is predicted to be a significant factor in intercity mode choice. Aggregated to the metropolitan level, interpolation of local travel activity is required, which may influence the validity of resulting demand models (Kobeissi & Habib, 2012). One possible reason for the lack of data is the issue regarding political jurisdictions as intercity travel information may require cooperation between neighbouring municipal, provincial and even federal agencies (Miller, 2004).

With the lack of available Canadian intercity travel data with local accessibility aspects, it is necessary to create a new survey framework to collect travel data for demand modelling. The main benefit of creating a new survey framework is the ability to dictate the information collected and to control the level of aggregation. Given the large focus on including local accessibility within an intercity trip, the relevant design aspects of the EcoTrain and Rand Corporation studies can be utilized and improved. Using a new survey instrument, data can be collected via a number of different methods. The more traditional methods have been; street intercepts, telephone interviews, or pen and paper surveys sent via mailing lists. Newer survey distribution methods have been through web-based survey instruments, which may require less time and cost resources per completed response.

Concurrently, developments in communication technology have led to increased household access to the internet. With the increased costs and declining effectiveness of traditional phone and mail based forms of data collection (Kaplowitz, Hadlock, & Levine, 2006), personal computing and mobile technology have become the new frontiers for researchers and planners to collect travel information. However, the effectiveness of large-scale internet based data collection has yet to be confirmed (Alvarez & VanBeselaere, 2005) given the rapid development
of this technology and the slow adaptation of researchers towards this new data collection medium.

The use of web-based survey instruments is a relatively new frontier of research-based data collection. The recent adaptation of web-based surveys by data collection projects such as the Canadian Census (Statistics Canada, 2011) and the most recent Transportation Tomorrow Survey (TTS) program (Data Management Group, 2012) is a result of the increased spread of internet availability. However, in most cases, a web-based survey has only been supplemented onto the existing survey program. For example, the 2011 TTS has a web-based survey component, but phone lists and advance letters were still required to inform the interviewee beforehand. In addition, the email only survey scenario at MSU had available email listings from university administration (Kaplowitz, Hadlock, & Levine, 2006). For web-based surveys, the current gap is the combined effectiveness of new survey recruitment techniques in the collection of data via web-based instruments.

In addition, web-based surveys have a limitation where the collection of complex data requires custom coding by the survey designer. If one wants to avoid custom coding a survey, the alternative to use third-party web survey service may limit the complexity of the survey. One such limitation may be the level of geographical aggregation collected. The dilemma to collect disaggregate data versus the ease of survey development is a hurdle that many researchers in all fields of studies deal with when surveys are required to be administered.

Initial expectations due to the introduction of a high speed rail line in the Quebec City – Windsor Corridor was in favour of considerable modal shift to high speed rail with significant induced travel. However, further insight into existing HSR feasibility studies and development of realistic travel scenarios in the proposed stated preference survey revealed the many barriers hindering the success of high speed rail in Canada. One such barrier is comparable local access and egress times compared with other non-automobile intercity travel modes, which could be a significant factor to the door-to-door travel time depending on the origin and destination. The purpose of this project is not to strictly assess modal shifts but to also determine which level-of-service attributes may have the most influence on intercity modal choice.
1.1 Objective

The goal of this project is to assess the potential modal shift of travellers with the introduction of a high speed rail line, the benefits of web-based survey instruments, and the feasibility of utilizing online social media for respondent recruitment. To achieve this goal, the completion of three key objectives is required.

The first objective is to understand the attributes that influence mode choice by estimating an analytical model that best represents choice. Additionally it should be possible to model the sensitivity of these significant attributes on a trip maker’s decision to travel via high speed rail rather than existing travel modes. A hybrid revealed preference (RP) and stated preference (SP) survey is proposed as a survey instrument to collect detailed data that can be used to statistically gauge an intercity trip makers’ perception of travel alternatives. The collected data should contain a representative sample of current intercity travelers in addition to individuals who have not made any intercity trips within the specified time period.

Secondly, the proposed web-based survey is designed as a dynamic survey tool, incorporating logical skips and background data queries to potentially reduce survey fatigue. The proposed project also attempts to use widely available software and web services to assess the advantages and disadvantages of creating a complex survey using tools that are easily and readily accessible to researchers.

The third objective is to assess the effectiveness of web-based survey methods as a possible replacement for traditional mail and phone based survey techniques. While many web-based surveys still depend on initial communication via letter or telephone, the goal is to propose an effective method of sample recruitment and data collection without using traditional sampling techniques, which may result in reducing overall survey costs.

1.2 Methods

To assess any potential modal shifts that would occur due to changing level-of-service attributes in intercity transportation modes, a novel data source is required beyond existing data sources. Further detailed in section 2.2, existing available public sources lack the geographical disaggregation potentially needed in understanding how local accessibility affects intercity modal choice. One of the more challenging aspects anticipated in the implementation of the
The proposed survey is to collect and analyze the effects of local travel accessibility on intercity mode choice. In addition, the proposed survey instrument would collect both revealed preference and stated preference data concurrently from a respondent. The following paragraphs outline the methodology of the proposed survey instrument as well as the intended data.

The instrument necessary to obtain the required data set is a combination of revealed preference (RP) information, stated preference (SP) information, and socioeconomic information. The study area of this project is mainly trips originating in the Greater Toronto Area (GTA). This geographic scope was limited to the GTA because of the emphasis of local transit accessibility, which increases the amount of work and computer time beyond a reasonable project scope for a larger geographic area. The revealed information is required to record previously experienced trips as well as daily local trip patterns. Examples of revealed information questions might include current modes of travel, travel time, travel costs, etc. The RP data would be used to characterize the selected sample as well as to define the most suitable subset of SP questions for the interviewee.

An additional aspect of required RP data is to determine whether or not the interviewee has previously made an intercity trip. While it is still important to assess the SP information from non-intercity trip makers, the combination of the RP and SP data can be jointly used to improve the estimation models of intercity mode choice.

In addition to the revealed preference data, a stated preference (SP) aspect is required. The SP section is intended to provide information of how respondents choose presented alternatives in a hypothetical situation. The benefits of a stated preference aspect within the survey instrument are the ability for SP surveys to statistically find the list and values of travel attributes that influence intercity mode choice.

A set of demographic information is also required to assess how accurately the smaller recruited sample of respondents describes a larger target population profile. While the sample frame is chosen to represent the larger population, the profile of collected results might not fully match the original sample frame. To minimize this bias, the demographic information of completed interviews should be verified with the original sample frame to assess statistical correlation, and then expanded to represent the population. The demographic information typically includes household location, age, gender, and income.
1.3 Thesis Outline

In the proposed project, a combined revealed preference – stated preference survey is designed and compiled into a web-based survey instrument. This thesis is laid out to emulate the individual steps required in designing and distributing the proposed RP-SP survey. Chapter 3 describes the conceptual design of the RP-SP survey. In Chapter 4, external data from several sources are combined together to output the stated preference tables. With a completed web-based RP-SP design and corresponding SP tables, Chapter 5 and 6 propose a set of procedures that is required to properly collect data from respondents using a web-based survey instrument. In Chapter 8, the collected data is used for model estimation.
Chapter 2
Literature Review

Over the past 50 years, there has been a doubling of Canada’s population; however, the amount of passenger travel has increased by more than five times (IBI Group, 2002). Parallel with the increased travel demand is the increasing density of urban centers as well as the connected infrastructure links. With the growth of urban centers, why do trip makers need to make intercity trips? While there have been many studies assessing why individuals make trip within an urban center, there is a lack of documentation for why individuals make intercity trips.

Studying intercity transportation is becoming increasingly more important as passenger travel demand continues to increase. Effective analysis of intercity travel demand could have impacts on the economic growth of major and minor urban centers as well as the infrastructure supporting travel between these urban centers.

2.1 Quebec City – Windsor Corridor

The Quebec-Windsor Corridor (QWC) stretches from Quebec City in Quebec to Windsor in Ontario. It influences the land area of approximately 100 km wide by 1,100 km long, this parcel of land contains approximately half of Canada’s population and 85% of Quebec and Ontario’s population (IBI Group, 2002). The map in Figure 2-1 below outlines the approximate area of the Quebec City – Windsor Corridor as well as the major metropolitan areas within the QWC. Due to existing concentrations of industry and population within this corridor, it can be considered the busiest and most important trade and transportation corridor in Canada (Patterson, Ewing, & Haider, 2007). The Montreal-Toronto section has become one of the busiest segments in the QWC in terms of both transportation and economic activities (Patterson, Ewing, & Haider, 2007).
Compared to the rest of Canada, the QWC is a highly localized economic center. Ontario typically accounts for 40% of the national total income and production. In 1994, (Polese, 2012) Ontario’s gross domestic product (GDP) was $302 billion, which is comparable to Belgium or Sweden. In addition, Ontario accounts for over 50% of Canadian production and roughly 65% of assets of Canada’s major financial institutions (Polese, 2012). Including Montreal’s urban region, which accounts for approximately 45% of Quebec’s income and production, the QWC is a booming center of economic activity. In total the Quebec City – Windsor corridor generates over 50% of Canada’s income and production and accounts for over 70% of its manufacturing employment (Polese, 2012).

With such a high amount of economic activity, transportation becomes an important linkage for the movement of both goods and people. EcoTrain’s report in section 2.3.1 further details the most recent dataset on traffic counts as well as the modal split of travelers within the Quebec City – Windsor Corridor.

Currently, the majority of the studies published regarding the QWC have been to suggest a new mode of transportation. The basis for this new transportation mode was to attempt at reducing the greenhouse gas (GHG) emissions related to automobile passenger travel. Within the transportation sector, passenger automobile accounts for 45% of the GHG production (IBI Group, 2002). Mass transit alternatives are found to be more efficient than automobiles, emitting 50% less GHGs than cars per passenger-km (Valli, 2010). There are also health and environmental costs of moving between cities. Ground-level ozone concentration has been
observed to be the highest within the QWC compared to the rest of Canada (Valli, 2010). Ontario Medical Association estimates that air pollution costs Ontarians over 1 billion dollars per year in hospitalizations, emergency room visits, and premature deaths (Valli, 2010).

Another argument regarding the implementation of a high speed rail system within the QWC is to reduce the cost of traffic congestion. In 2006, Transportation Canada’s urban traffic congestion report determined congestion was costing Canadians between 2.3 to 3.7 billion dollars annually (Valli, 2010). While these costs are not entirely on intercity travel, the increased congestion that would occur from increased intercity travel under business-as-usual scenarios would increase the stated congestion costs.

### 2.2 Intercity Travel Demand

Intercity travel is defined as a trip that passes through the boundaries of an urban center. Typically an intercity trip would start and end within an urban center. Intercity travel demand would be the frequency of trips made between urban centers as well as the modes of travel used. While intercity travel demand models have been published since the 1960s, these models have not been improved as rapidly as the more urban center specific models. Eric Miller’s paper (Miller, 2004) hypothesized reasons for the slow development of intercity models due to:

- Fewer intercity travel corridors of interest to policy makers in comparison to urban regions. In Canada, the major corridors would be the Quebec City – Windsor Corridor, Calgary – Edmonton Corridor, and the Vancouver – Seattle Corridor.
- Jurisdiction of intercity corridors cannot be defined by one single planning agency. The EcoTrain report in section 2.3.1 is comprised of both federal and provincial transportation agencies as well as a consortium of international consulting firms to produce this one feasibility report (EcoTrain, 2011).
- Larger private sector stake resulting in proprietary information. Neither the stated preference survey nor the dataset from the survey are publicly available for further analysis, which may hinder progressive development of demand models.
- Open-ended definition of study area makes it hard to collect meaningful data. With such a large geographical area within a transportation corridor, it is difficult to balance between a workable scope and comprehensive results.
• Existing data available for research purposes are limited.
• Governments and public agencies unwilling to invest in long-term research.

Also according to Miller, almost all intercity travel demand models consists of three stages; a direct demand model to predict the number of trips for each origin-destination pair, a mode split model to determine the travel mode for each trip, and a trip assignment model to yield link flows. This process is similar to a conventional four-step urban transportation modelling system with the exception that the direct demand model combines trip generation and trip distribution into one equation (Miller, 2004). The following sections below summarize previous intercity travel demand models in Canada and other parts of the world.

Application of models such as the logit model to assess intercity travel demand has not been as successful as the urban center counterparts. Aside from the lack of suitable data, there are other fundamental reasons. Compared to daily local journeys (trips to work, school, shopping), intercity trips are made with longer intervals than urban trips (Sonesson, 2001). Using the logit model, there is difficulty including a trip maker’s decision to visit another city as this decision was based on the previous trips (and even future trips) to the same city. Unless daily intercity trips are made, a longer time period of trip information collection is required (Sonesson, 2001). The following sections summarize Canadian-based intercity demand models as well as international intercity demand models.

### 2.2.1 Existing Canadian Travel Demand Models

An empirical study in using disaggregate choice model of intercity travel demand in the Quebec City – Windsor Corridor was published in 1988 by Ridout and Miller (Ridout & Miller, 1982). The main data set used was the 1969 Canadian Transport Commission (CTC) survey of mode choice in the QWC. The major weaknesses in the data set was the lack of automobile choice, which limited the accuracy of the demand forecasting logit model as automobile comprises a large market share in intercity travel. While the 1980 Canadian Travel Survey (CTS) was available as a data set alternative, the aggregation of trip origins and destinations to the Census Metropolitan Area (CMA) would not provide relevant trip access and egress information (Ridout & Miller, 1982). The resulting models presented by Ridout and Miller were all calibrated using a standard maximum likelihood logit estimation procedure (Ridout & Miller, 1982). Three models were estimated with each model representing different market segmentation (business, pleasure,
or personal). For the business market, it was observed that access was a generic attribute while egress was alternative specific. For non-business and non-personal intercity trips, access was estimated as an alternative specific attribute and egress was generic. For personal intercity trips, only access, as a generic attribute, was estimated to be statistically significant to mode choice (Ridout & Miller, 1982). Ridout and Miller experienced difficulty in estimating the access and egress term, which may be attributed to the use of access/egress distance instead of cost and time. Socioeconomic variables were found to have minimal explanatory function in intercity mode choice (Ridout & Miller, 1982). One explanation for this lack of socioeconomic explanatory variables may be the nature of socioeconomic characteristics contributing to trip generation rather than mode choice.

Another disaggregate mode choice model of Canadian intercity passenger travel was published in 1988 by Wilson, Damodaran, and Innes. The goal of this paper was to identify significant level-of-service factors that influenced intercity mode choice for further development of disaggregate models and for policy impact analysis (Wilson, Damodaran, & Innes, 1990). Using the 1985 Canadian Travel Survey (CTS), the data available for model estimation included socioeconomic information as well as trip information (for all trips greater than 80km) from a sample size of 19,000 households within Canada (Wilson, Damodaran, & Innes, 1990). A limitation in the CTS data was the reporting of only trip distances and not trip times, which required estimation of travel times using a standardized function of average travel speed (Wilson, Damodaran, & Innes, 1990). The data was segmented into business and nonbusiness trips as well as two geographic boundaries. A multinomial logit form was used for the estimated models and four mode choice alternatives were considered in the model; air, auto, bus, and rail. The estimated models had reasonable goodness-of-fit measures with the models favoring travel by auto and air more than bus and rail (Wilson, Damodaran, & Innes, 1990).

To potentially overcome the independence of irrelevant alternatives (IIA) property in the previously used multinomial logit models, Bhat proposed to use a heteroscedastic extreme value (HEV) model to estimate mode choice (Bhat, 1995). The IIA restructure is avoided by allowing the random components of utilities to have unequal scale parameters. The HEV model should allow for cross-elasticity among alternatives compared to a nested logit model and require less computational complexity compared to the multinomial probit model (Bhat, 1995). To test the application of the HEV model, the 1989 Rail Passenger Review from VIA Rail was used. This
data set included both socioeconomic information and trip characteristics within the Quebec City – Windsor Corridor (Bhat, 1995). The main focus was on paid business travel in the QWC and confined between the air, rail, and auto travel modes. Dropping the bus alternative was justified as bus travel via paid business purposes had less than 1% market share (Bhat, 1995). Five different models were estimated; a multinomial logit model, three nested logit models, and the heteroscedastic extreme value model. From model estimation, the nested logit structure was not significantly better than the multinomial logit models or was inconsistent with utility maximization principles (Bhat, 1995). Compared to both MNL and NL, the HEV model was able to predict smaller changes in level-of-service changes, which may point to an improvement of the HEV model over the commonly used MNL and NL model formulations (Bhat, 1995).

It is often hypothesized that an individual’s responsiveness to level-of-service variables affects that individual’s mode choice (Bhat, 1998). In this paper, Bhat accommodates variations in this responsiveness within a multinomial logit based model. Monte Carlo simulation techniques were also incorporated to approximate the choice probabilities, which are a technique that has been used in empirical applications in the economics field and relatively new to transportation researchers (Bhat, 1998). The 1989 Rail Passenger Review data set from VIA Rail is used once again to develop the travel demand models. Once again, weekday business-based market segment was selected and recorded trips by bus were omitted from the data set due to a small percentage of market share (Bhat, 1998). Three models were estimated in this paper; a multinomial logit model, a fixed-coefficient logit model, and a random coefficients logit model. Level-of-service variables included service frequency, total travel cost, in-vehicle travel time and out-of-vehicle travel time along with socioeconomic variables; income, sex, travel group size, and large city indicator (Bhat, 1998). From model estimation, both the fixed-coefficient logit and random coefficient logit models showed differences in sensitivity of level-of-service variables based on the socioeconomic characteristics of the trip maker, rejecting the response homogeneity assumption of the MNL model (Bhat, 1998). Bhat shows that not accounting for variations in responsiveness across individuals may lead to inappropriate estimation of mode choice, which may affect policy decisions.

Another alternative to address the independence of irrelevant alternatives property of multinomial logit models was Koppelman and Wen’s use of the paired combinatorial logit (PCL) model. PCL model relaxes the IIA restricture by allowing different covariances for each pair of
alternatives, which enables the estimation of differential competitive relationships between each alternative pair (Koppelman & Wen, 2000). The 1989 Rail Passenger Review data set from VIA Rail was used to estimate the demand for high speed rail in the Toronto to Montreal corridor. The data set was filtered to include the 2,769 individuals who have air, rail, and auto modes available (Koppelman & Wen, 2000). From model estimation, paired combinatorial logit model allows for utility correlation and substitution between each pair of alternatives. Compared to both multinomial and nested logit models, PCL produced better estimations and has application in producing forecasts based on the alternative under study as well as alternatives from which those travelers are likely to be drawn (Koppelman & Wen, 2000).

Building on the paired combinatorial logit model, Wen and Koppelman introduced the generalized nested logit (GNL) model in a 2001 paper. The GNL model accommodates differential cross-elasticity of pairs of alternatives and provides higher degree of flexibility in the estimation of substitution or cross-elasticity compared to previously developed generalized extreme value (GEV) models (Wen & Koppelman, 2001). This GNL model also includes the paired combinatorial logit (PCL), two level nested logit (NL), and cross-nested logit (CNL) models as special cases as well as the product differentiation PD model (Wen & Koppelman, 2001). Similar to the previous model estimations, the 1989 Rail Passenger Review data set from VIA Rail was used (Wen & Koppelman, 2001). From model estimation, it was observed that the GNL model adds useful flexibility to existing models under the GEV model family. The GNL model provides a more flexible structure to estimate differential cross-elasticities among pairs of alternatives and provides a unifying structure (Wen & Koppelman, 2001). One potential advantage of the GNL model is the ability to explore cross-elasticity structures without necessarily estimating a large number of NL models (Wen & Koppelman, 2001).

2.2.2 Existing International Travel Demand Models

In Ghoneim and Sargious’ 1987 paper on disaggregate mode choice models in intercity passenger rail; a review of existing research indicated that disaggregation is statistically and behaviorally necessary to model human travel behavior (Ghoneim & Sargious, 1987). In differentiating between aggregate and disaggregate mode choice models, there are three key data attributes that are required in order to calibrate a disaggregate model. The first key attribute is comprehensive coverage of all intercity travel; modes, purposes, duration, and characteristics of
the trip and trip maker. The second key attribute is consistency in data collection methodology and definition of variables. The last key attribute is the design of survey forms that provides individual-level trip information to induce spatial disaggregation (Ghoneim & Sargious, 1987).

The first reviewed study was a 1976 paper by Stopher and Prashker to test the adequacy of disaggregate behavioral demand modelling in the context of intercity travel. Using the American 1972 National Travel Survey, a multinomial logit mode choice model of four intercity travel modes (air, rail, bus, and auto) was calibrated for four major city pairs in the United States (Stopher & Prashker, 1976). The conclusion of this paper confirmed the applicability of disaggregate approach to intercity mode modelling but also pointed out deficiencies in the National Travel Survey data, which produced counter-intuitive results in sensitivity and policy analysis test (Stopher & Prashker, 1976). A second paper reviewed by Ghoneim was a 1982 paper by Grayson to develop policy sensitive models for intercity mode choice. Grayson’s paper used an updated 1977 National Travel Survey and found the absence of auto travel costs, frequency of common carrier modes, and unreliable travel times. The resulting intercity mode choice model was satisfactory to assess policy sensitivity but precluded differentiation between longer and shorter trips (Grayson, 1981).

Unlike the two earlier models, Stephanedes and Kumar developed a mode choice model for intercity business travel from a fully disaggregate data base, which accounts for the buying power of a trip maker as well as the trip length. This disaggregate data was sourced from the 1982 Twin Cities-Duluth Trip survey (Stephanedes, Kumar, & Padmanabhan, 1984). In this survey, airplane, bus, and auto trips were presented as choice alternatives. In the data set, access time, distance, and cost for non-chosen alternatives were estimated to ensure full data disaggregation. From data collection, 90 observations were deemed usable for analysis. The estimation results suggest that out-of-pocket travel costs was consistently the most significant parameter in intercity mode choice. In addition, a purchasing power attribute was created by computing the travel cost by income (Stephanedes, Kumar, & Padmanabhan, 1984). This purchasing power attribute was highly significant and in agreement with most work-trip mode choice models. Three models were estimated with the collected data and each model considered; in-vehicle travel time, out-of-vehicle travel time, out-of-pocket travel cost, and household income. Results of model estimation concluded that the proposed model, using the disaggregate
data set, performs up to 66 percent better than previous models and is useful to assess policy impacts from common carriers (Stephanedes, Kumar, & Padmanabhan, 1984).

In the deregulation of the U.S. commercial air transportation system in 1978, the increased public concerns over safety and rising fares (Morrison & Winston, 1989). These concerns motivated Morrison and Winston to assess possible improvements to the air travel system by creating models to better understand the effects of policy changes such as; mergers, pricing, investments, safety, and traffic volumes. A disaggregate airline carrier choice model was estimated using a random sample of air passenger round-trips collected in 1983 (Morrison & Winston, 1989). Using an individual’s choice of carrier and routing, alternative carriers and routes were generated. With the collected and generated data, the air traveler multinomial logit choice model was estimated using level-of-service attributes such as; average far, travel time, schedule delay, transfer time, number of complaints, percentage of flights on time, and size of air carrier company (Morrison & Winston, 1989). From model estimation, it was also possible to infer on the traveler’s value of travel time, transfer time, and schedule delay in dollars per hour.

Mode choice analysis is the third step within the classical four-step transportation planning progress. Previous research into mode choice has revealed that the multinomial logit model and variations of logit models are beneficial to determining mode choice. Mandel, Gaudry, and Rothengatter further enrich the logit mode choice specification by including mode attributes, socioeconomic variables, and trip purpose characteristics to improve model quality. Additionally, Box-Cox transformations applied to model attributes yield better results than linear logit models (Mandel, Gaudry, & Rothengatter, 1997). This Box-Cox transformation bypasses most of the constraints of linear logit specification including; equal cross elasticities of demand, exclusion of complementarity among alternatives, symmetric response curve to changes in service characteristics, and under-identified coefficients for the constants and generic variables. Using the Box-Cox device, it is possible to continuously explore many ad-hoc specific transforms; however, this efficiency comes at the cost of a unique maximum if more than one variable is transformed in a non-linear form (Mandel, Gaudry, & Rothengatter, 1997). To test the application of the Box-Cox device, demand models were estimated using trip information from the 1979/1980 KONTIFERN database from the German Ministry of Transportation. This database includes origin-destination attributes, trip purpose attributes, and socioeconomic variables of long distance passenger travel (Mandel, Gaudry, & Rothengatter, 1997). Four model
series were estimated and ranged between simplest to most complex use of database variables. From model estimation, it was observed that nonlinearity was better than the linear logit model from a theoretical and practical point of view. It was found that linear forms tend to over-predict market shares for short distances and under-predict market shares for longer distances. The influence of high speed rail links on consumer behavior was found to be stronger in the nonlinear case compared to the linear one (Mandel, Gaudry, & Rothengatter, 1997).

Many of the previous intercity demand models have been estimated with revealed preference (RP) data. In the case where non-chosen alternatives were incorporated into the model, such as Morrison and Winston’s 1989 study on airline carrier choice, non-chosen alternative attributes were generated after the fact. A further step in demand modelling may be the use of stated preference (SP) data to model demand. In Hensher, Louviere, and Swait’s 1999 paper on combining sources of preference data, a combined RP-SP data set is used to estimate the market for a proposed high speed rail system in the Sydney-Canberra Corridor in Australia (Hensher, Louviere, & Swait, 1999). The addition of SP data is beneficial when; modelling demand for new products with no RP history, key RP explanatory variables have little to no variability, and attributes previously not present may now influence choice (Hensher, Louviere, & Swait, 1999). Data was collected from a random sample of individuals who made freight shipper choice decisions. The survey instrument collected firm characteristics, RP data of the respondent’s most recent carrier choice, and SP choice data of 16 binary choice scenarios. For the SP choice questions, 80 scenarios were generated using an orthogonal main effects design and blocked off into five sets of 16 choice scenarios each (Hensher, Louviere, & Swait, 1999). From model estimation, several models were calibrated, each with a different functional form. The first model is a RP multinomial logit model (MNL) with homoscedastic error variances, which yielded statistically significant coefficients and correct signs. The second model is a RP parameterized heteroscedastic multinomial logit model (PHMNL), which relaxes the homoscedasticity restriction. From these two models, it was concluded that the RP data source exhibits evidence of heteroscedastic error structure. The third model is a SP homoscedastic MNL model with all parameters statistically significant and yielding the right sign. The fourth model is a SP PHMNL models and estimated statistically equal variances for all alternatives, which suggests that unobserved characteristics do not induce different levels of variance in the SP data. From model
estimation, it was observed that multiple data sources can be used to better estimate choice models especially in estimating demand of future markets (Hensher, Louviere, & Swait, 1999).

Alongside the potential modal shifts that may occur with the introduction of a new travel mode are induced travels, which are the additional trips that would not have been made if the new travel mode was not available. Previous studies have shown that travel demand growth within a specific corridor tends to be greater than the expected travel growth after new capacity is added (Yao & Morikawa, 2005). Yao and Morikawa’s 2005 paper utilizes a combined RP-SP data set to model choice demand and potential induced travel in the Tokyo-Nagoya-Osaka (TNO) corridor given the introduction of a 500km/h high speed rail line in the corridor. The questionnaire was distributed to railway users from six of Japan’s metropolitan areas where respondents were asked to record his or her last trip as well as to complete a SP survey. In the SP survey, eight choice scenarios were presented with varying attributes of a hypothetical high speed rail mode choice (Yao & Morikawa, 2005). With the collected data, an integrated mode/route choice model was estimated, which follows a nested structure. From model estimation, all parameters are statistically significant and consistent with hypothesized signs on utility. The explanatory variables were; travel cost, line-haul time, access/egress time, and service frequency for both business and non-business market segments. In addition, the model estimated 14.5% induced travel demand within the TNO corridor at the expected completion of the high speed rail service (Yao & Morikawa, 2005).

Although long distance trips (over 50 miles one-way) accounts for less than 2% of the total number of trips made by British residents travelling within Great Britain, these trips account for 30% of the distance traveled (Dargay & Clark, 2012). Dargay and Clark assess the effects of socioeconomic, demographic, and geographic factors of intercity travel in a 2011 paper. The National Travel Survey of Great Britain residents from 1995 to 2006 were used as the main dataset. Econometric analysis of the dataset was through linear regression and included suitable explanatory variables. The results of model estimation concluded that long distance travel is strongly related to income (Dargay & Clark, 2012).

### 2.3 High Speed Rail

In global context, high speed rail (HSR) is not a transportation mode with a set specifications listing. Instead, there are a number of existing technologies that could be considered a high speed
rail system. Based on the feasibility from EcoTrain, two representative technologies will be discussed with one technology being chosen for use with the stated preference survey.

The first type is a modern high-speed diesel fueled technology with a minimum operating speed of 200 km/h or more (F200+). For this type of HSR, little infrastructure improvements are required if existing rail alignments are used. Because it is diesel based, the maximum installed power is around 4,500 kW, which limits the maximum speed of a typical 200m long train to 200 km/h (EcoTrain, 2011). From an environmental perspective, the use of diesel does contribute to direct emission of greenhouse gasses and other air contaminants.

The second specification being considered is a modern high-speed electric technology, with a minimum operating speed of 300 km/h or more (E300+). As this technology is reliant on electricity, higher initial infrastructure costs are required to construct electrified lines and distribution systems. Given a higher initial cost, the benefit of electrified lines is a higher traction power of 10,000 kW, allowing the train to run at maximum speeds of 300 km/h (EcoTrain, 2011). Electric-based technologies also allow the train to recover braking energy as an additional energy source (EcoTrain, 2011). Compared to the fossil fuel dependent diesel technology, if electricity is produced from a cleaner source of energy, there are lower environmental impacts in both the emitted greenhouse gas and air contaminants.

Assuming that all other costs of implementing a high speed rail system is the same between the F200+ and E300+ alternatives, the electric-based E300+ option is chosen as the implemented HSR for the stated preference aspect of the proposed survey design.

Many government funded feasibility studies have been done over the years regarding the implementation of a high speed rail system in the Quebec-Windsor corridor. Many of these studies have concluded that high speed rail is feasible as a long-term project. These studies include the 1995 Ontario High Speed Rail Project Study, the 2002 Case Study of the Quebec City – Windsor Corridor, the 2006 EcoTrain Updated Feasibility Study. However, there have been many hurdles delaying its implementation. Political barriers include fear that the government would not recover from the cost of an estimated 25 billion dollar construction project (Valli, 2010); however it should be noted that 2001 government subsidies for VIA Rail was in the range of 170 million dollars (Transport Canada, 2003).
2.3.1 Ecotrain Feasibility Study

The most recent feasibility study regarding the implementation of high speed rail service in the Quebec City – Windsor Corridor was completed in 2010 by EcoTrain, which is a group of consulting firms lead by Dessau and comprised of Deutsche Bahn International (DBI), KPMG, MMM Group, and Wilbur Smith Associates (WSA). This report is a comprehensive update of a previous 1995 feasibility report. The main purpose of this updated feasibility study was to consider the elements of; technologies, routes, stations, service, travel demand forecasting, investment costs, operations and maintenance costs, environmental and social aspects, implementation schedule and options, financial and economic analysis, impacts on the transportation system, review of similar HSR projects, and review of transportation policies in Europe.

The technologies, routes, and service elements are referenced throughout the proposed project and are used as baseline values of the high speed rail travel attributes in the proposed stated preference survey. While these elements are not the defined specifications of a high speed rail service, the speculated attributes are based on considerations with constructability, existing services, and technical limitations. For example, it was speculated that a high speed rail service could share tracks with commuter rail networks in Montreal (AMT) and Toronto (GO Transit). This speculation outlines possible feasibilities despite the fact that further discussion and design of the potential to share tracks has yet to be publically detailed.

The EcoTrain feasibility report is a highly technical report that outlines social, environmental, and technical impacts of introducing a high speed rail service. While all of these aspects are important to understanding the overall feasibility of introducing a high speed rail service, emphasis should be placed on the demand forecasting aspect of the EcoTrain feasibility report. This main objective of updating demand forecasting was to “…update OD (auto) intercity trip estimates using new survey data…” and update “…ridership diversion rates using new demand modelling calibrated against a stated-preference survey that must reflect observed (or revealed) modal choices…”. The similarity of this element of the EcoTrain feasibility report to the proposed project objectives is beneficial as reference material.

Focus of the demand forecasting section of the EcoTrain report was placed on; updating non-automobile travel data conducting a license plate survey to estimate auto travel in the corridor,
and implementing a stated preference survey to develop a new set of mode choice models. From these three (3) tasks, forecasting potential ridership on a high speed rail service would be estimated by:

- Determining total trips made in the corridor for all existing travel modes, which represents the total travel market that could potentially divert to a new high speed rail service.
- Estimating the market share that would be captured by high speed rail based on travel cost (fare), travel time, and other service attributes of all potential travel modes.
- Estimate the number of induced trips, for example the number of trips that would not be made if high speed rail was not available.

One of the expected difficulties in estimating demand is the different characteristics of travel based on market segmentation. For example, an individual making a business trip may be less sensitive to travel cost and more sensitive to travel time compared to recreational or social travelers. This difference in market segmentation would potentially require individual demand forecasting models, which may require larger data collection sets.

As no new Origin Destination (OD) surveys were conducted for public modes within the project scope of EcoTrain, the updated public modes OD ridership data was taken from existing travel data for 2006 provided by Transport Canada. Table 2-1 below is a summary of the 2006 trip values for both business and non-business market segments for the air, bus, and rail intercity transport modes within the Quebec City – Windsor Corridor (QWC).

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th>Rail</th>
<th>Air</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major Markets</td>
<td>438,000</td>
<td>693,000</td>
<td>1,871,000</td>
<td>3,002,000</td>
</tr>
<tr>
<td>Other Markets</td>
<td>181,000</td>
<td>215,000</td>
<td>326,000</td>
<td>722,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>618,000</td>
<td>908,000</td>
<td>2,198,000</td>
<td>3,724,000</td>
</tr>
<tr>
<td><strong>Non-Business</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major Markets</td>
<td>1,775,000</td>
<td>1,342,000</td>
<td>320,000</td>
<td>3,437,000</td>
</tr>
<tr>
<td>Other Markets</td>
<td>1,684,000</td>
<td>965,000</td>
<td>84,000</td>
<td>2,733,000</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3,459,000</td>
<td>2,307,000</td>
<td>403,000</td>
<td>6,169,000</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td>4,078,000</td>
<td>3,215,000</td>
<td>2,601,000</td>
<td>9,894,000</td>
</tr>
</tbody>
</table>
Based on the large modal share of auto travel within the QWC (91% in 1992), a more critical method of data collection was completed to assess auto travel. For one week day and one week end of a period of 12 hours each, data was collected at six (6) locations which were deemed most critical to capturing intercity auto travel in the QWC. The method of data collection was to capture license plate information for all Private Occupancy Vehicles (automobiles, SUVs and pickup trucks) travelling in both directions using special Automated License Plate Recognition cameras. In addition a directional mechanical vehicle classification count was also conducted at the six (6) survey locations to count the actual number and class of vehicles. At the conclusion of the license plate survey, 93,205 license plates were collected from Ontario stations and 54,481 license plates were collected from Quebec stations for a total of 147,686 license plates.

A random sample of captured license plates was forwarded to SAAQ and MTO to obtain names and addresses of vehicle owners in order to send him or her a questionnaire regarding the intercity trip. This survey could be filled out by pen and paper or via a web-based survey and asked only revealed preference questions regarding the individual’s intercity trip information and basic socioeconomic information. A copy of this survey is attached in Appendix A. A total of 11,415 valid responses were collected through both letter and web-based surveys.

The full results of the License Plate survey and subsequent data analysis on automobile based travel in the Quebec City – Windsor Corridor is available in the EcoTrain Demand Forecast report. A summary of the 2009 License Plate results is illustrated in Table 2-2 below.

<table>
<thead>
<tr>
<th>City Pair</th>
<th>City Pair</th>
<th>Number of Surveys</th>
<th>Expanded Number of Trips</th>
<th>Maximum Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quebec</td>
<td>Montreal</td>
<td>1,116</td>
<td>6,648,000</td>
<td>±7%</td>
</tr>
<tr>
<td>Quebec</td>
<td>Ottawa</td>
<td>88</td>
<td>427,000</td>
<td>±31%</td>
</tr>
<tr>
<td>Quebec</td>
<td>Toronto</td>
<td>27</td>
<td>67,000</td>
<td>±75%</td>
</tr>
<tr>
<td>Montreal</td>
<td>Ottawa</td>
<td>726</td>
<td>3,875,000</td>
<td>±7%</td>
</tr>
<tr>
<td>Montreal</td>
<td>Kingston</td>
<td>54</td>
<td>305,000</td>
<td>±33%</td>
</tr>
<tr>
<td>Montreal</td>
<td>Toronto</td>
<td>367</td>
<td>1,563,000</td>
<td>±16%</td>
</tr>
<tr>
<td>Montreal</td>
<td>London</td>
<td>17</td>
<td>59,000</td>
<td>±89%</td>
</tr>
<tr>
<td>Montreal</td>
<td>Windsor</td>
<td>4</td>
<td>28,000</td>
<td>±100%</td>
</tr>
<tr>
<td>Ottawa</td>
<td>Kingston</td>
<td>67</td>
<td>1,380,000</td>
<td>±24%</td>
</tr>
<tr>
<td>Ottawa</td>
<td>Toronto</td>
<td>315</td>
<td>2,656,000</td>
<td>±14%</td>
</tr>
<tr>
<td>Ottawa</td>
<td>London</td>
<td>31</td>
<td>229,000</td>
<td>±56%</td>
</tr>
</tbody>
</table>
Table 2-2 – 2009 License Plate Survey Results (EcoTrain, 2011)

<table>
<thead>
<tr>
<th>City</th>
<th>City</th>
<th>Count</th>
<th>Population</th>
<th>± Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ottawa</td>
<td>Windsor</td>
<td>16</td>
<td>115,000</td>
<td>±89%</td>
</tr>
<tr>
<td>Kingston</td>
<td>Toronto</td>
<td>145</td>
<td>1,859,000</td>
<td>±19%</td>
</tr>
<tr>
<td>Kingston</td>
<td>London</td>
<td>14</td>
<td>160,000</td>
<td>±71%</td>
</tr>
<tr>
<td>Toronto</td>
<td>London</td>
<td>227</td>
<td>5,515,000</td>
<td>±11%</td>
</tr>
<tr>
<td>Toronto</td>
<td>Windsor</td>
<td>59</td>
<td>1,388,000</td>
<td>±28%</td>
</tr>
</tbody>
</table>

A stated preference survey was designed and administered by EcoTrain to collect information from individuals who were making or had made a trip within the QWC that was within the feasible high speed rail routes.

Data was collected by intercept at Montreal Trudeau International Airport and central bus/VIA rail stations in Montreal, Toronto, Ottawa, and Quebec City. Additionally, data was collected at convention centers and central libraries in the four cities. Some respondents from the license plate survey were also contacted to complete the stated preference survey. The objective was to collect data from current auto, bus, rail, and air travelers in the QWC. The following Table 2-3 is a breakdown of the survey respondents.

Table 2-3 – Stated Preference Survey Results (EcoTrain, 2011)

<table>
<thead>
<tr>
<th>Survey Type / Province</th>
<th>Total Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontario License Plate Survey</td>
<td>318</td>
</tr>
<tr>
<td>Ontario Field Intercept Location</td>
<td>384</td>
</tr>
<tr>
<td>Sub-Total Ontario</td>
<td>702</td>
</tr>
<tr>
<td>Quebec License Plate Survey</td>
<td>392</td>
</tr>
<tr>
<td>Quebec Field Intercept Location</td>
<td>459</td>
</tr>
<tr>
<td>Montreal Trudeau Airport</td>
<td>149</td>
</tr>
<tr>
<td>Sub-Total Quebec</td>
<td>1,000</td>
</tr>
<tr>
<td>Grand Total</td>
<td>1,702</td>
</tr>
</tbody>
</table>

For each respondent, eight stated preference scenarios were presented. If the intercept survey was during an intercity trip, then the scenario of the stated preference questions were related to the respondent’s currently ongoing trip. For each scenario, three (3) alternatives were provided:

1. Current travel mode
2. Travel by a hypothetical high speed rail mode
3. Travel by a third randomly selected mode which is either auto, bus, rail, or air but could not be a repetition of the currently traveled mode
In each scenario, the respondent essentially ranks the three modes by choosing the mode most likely chosen as well as the mode least likely chosen. A set of attributes were shown which describes the details of the trip for each of the three (3) travel modes. These attribute included; travel time, parking costs, and fare. The values of the attributes were derived from an orthogonal experimental design. The below Figure 2-2 is a sample of one scenario of the stated preference survey.

![Sample SP Scenario in EcoTrain Survey](EcoTrain, 2011)

With the data collected, the collection of both revealed preference and stated preference data were used to estimate a nested logit (NL) choice model for both business trips and non-business trips. For both segments, the utility equation structures were tested using variables in the SP survey as well as RP trip characteristics and demographic variables. To test potential interactions with intercity travel costs and travel time, other variables were introduced into the model estimation process one at a time.

Weighting of the data set was based on intracity mode-specific trip totals from 2006 data. Table 2-4 below is the weights used for both market segments for the different primary modes.

<table>
<thead>
<tr>
<th>Segment</th>
<th>Primary Mode</th>
<th>Province</th>
<th>Weight Applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>Auto</td>
<td>Quebec</td>
<td>1.6718</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ontario</td>
<td>1.6789</td>
</tr>
</tbody>
</table>
Multiple nesting structures were considered during the model estimation process. Rather than a behavioral tree based approach, an economic tree was used. The resulting model shows that alternatives within a nest are better substitutes for each other compared to alternative not within the nest, meaning that nested alternatives have higher cross-elasticities. The result of EcoTrain’s nesting structure is detailed in Appendix B.

An elasticity study was also completed to analyze the elasticity of mode choice models to get an understanding of the relative impact on mode choice due to level of service changes. Cross-elasticity provides the impact a change has on other modes. The following series of tables are the resulting elasticity of business and non-business models for Quebec to Montreal travel. Direct elasticities are along the diagonals and cross-elasticities are found in off-diagonals.
### Table 2-6 – Fare Elasticity of Non-Business Model for Quebec to Montreal Segment (EcoTrain, 2011)

<table>
<thead>
<tr>
<th>LOS-Fare</th>
<th>Air</th>
<th>Automobile</th>
<th>VIA Rail</th>
<th>Bus</th>
<th>HSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air</td>
<td>-0.2534</td>
<td>0.1935</td>
<td>0.0035</td>
<td>0.0122</td>
<td>0.0441</td>
</tr>
<tr>
<td>Automobile</td>
<td>0.0115</td>
<td>-0.2235</td>
<td>0.8451</td>
<td>0.8451</td>
<td>0.1434</td>
</tr>
<tr>
<td>VIA Rail</td>
<td>0.0141</td>
<td>0.0188</td>
<td>-1.296</td>
<td>0.0188</td>
<td>0.1764</td>
</tr>
<tr>
<td>Bus</td>
<td>0.0132</td>
<td>0.0616</td>
<td>0.0616</td>
<td>-1.1714</td>
<td>0.1654</td>
</tr>
<tr>
<td>HSR</td>
<td>0.0074</td>
<td>0.4056</td>
<td>0.0073</td>
<td>0.0256</td>
<td>-0.446</td>
</tr>
</tbody>
</table>

### Table 2-7 – In Vehicle Travel Time of Business Model for Quebec to Montreal Segment (EcoTrain, 2011)

<table>
<thead>
<tr>
<th>LOS-Fare</th>
<th>Air</th>
<th>Automobile</th>
<th>VIA Rail</th>
<th>Bus</th>
<th>HSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air</td>
<td>-0.308</td>
<td>0.1317</td>
<td>0.0028</td>
<td>0.0095</td>
<td>0.0074</td>
</tr>
<tr>
<td>Automobile</td>
<td>0.0012</td>
<td>-0.0672</td>
<td>0.0091</td>
<td>0.0309</td>
<td>0.0261</td>
</tr>
<tr>
<td>VIA Rail</td>
<td>0.0014</td>
<td>0.5267</td>
<td>-1.196</td>
<td>0.1871</td>
<td>0.0322</td>
</tr>
<tr>
<td>Bus</td>
<td>0.0014</td>
<td>0.4938</td>
<td>0.5963</td>
<td>-0.7007</td>
<td>0.0302</td>
</tr>
<tr>
<td>HSR</td>
<td>0.3439</td>
<td>0.276</td>
<td>0.0059</td>
<td>0.0199</td>
<td>-0.3172</td>
</tr>
</tbody>
</table>

### Table 2-8 – In Vehicle Travel Time of Non-Business Model for Quebec to Montreal Segment (EcoTrain, 2011)

2.3.2 RAND Corporation Feasibility Study

A study was commissioned by the UK Department for Transport to develop a model to predict passenger demand for long-distance travel, which would be used to assess policy changes such as the implementation of high speed rail. In the context of the study, long distance trips are defined as one-way journeys over 50 miles (80.5 kilometers). A two phase study was started in 2008 to develop a multi-modal model for long-distance travel. The second phase of the study was to design and implement a stated preference study to assess possible modal shifts due to the introduction of high speed rail service.

The stated preference survey is based on a high speed rail service linking London and Scotland via the west and east coast with a number of intermediate stops. This survey targeted existing trip makers within this corridor to strengthen better validate the collected responses. Respondents were collected by:

- Large-scale random sample household survey for rail and car travelers. Respondents were contacted using phone, mail, email, and internet collection methods.
- Computer assisted personal interviewing (CAPI) surveys for rail travelers undergoing train trips.
• CAPI surveys for air travelers at airports.
• Additional sample telephone list to obtain a representative British population from initial concerns that the original sample would not be met through the household survey.

The resulting data collection efforts collected a total of 3,045 responses. Table 2-9 and Table 2-10 below reveal both the distribution of data collection methods as well as the breakdown of collected data by mode and trip purpose.

<table>
<thead>
<tr>
<th>Car</th>
<th>Rail</th>
<th>Air</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>838</td>
<td>288</td>
<td></td>
<td>1,126</td>
</tr>
<tr>
<td>165</td>
<td>30</td>
<td></td>
<td>195</td>
</tr>
<tr>
<td>705</td>
<td></td>
<td></td>
<td>705</td>
</tr>
<tr>
<td>1,019</td>
<td></td>
<td></td>
<td>1,019</td>
</tr>
<tr>
<td>1,003</td>
<td>1,023</td>
<td>1,019</td>
<td>3,045</td>
</tr>
</tbody>
</table>

Table 2-9 – Breakdown of SP Interviews by Mode and Survey Approach (Burge, Kim, & Rohr, 2011)

<table>
<thead>
<tr>
<th>Employer’s business</th>
<th>Car</th>
<th>Rail</th>
<th>Air</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employer’s business</td>
<td>262</td>
<td>433</td>
<td>631</td>
<td>1,326</td>
</tr>
<tr>
<td>Commute</td>
<td>25</td>
<td>75</td>
<td>-</td>
<td>100</td>
</tr>
<tr>
<td>Visiting friends or relatives</td>
<td>716</td>
<td>515</td>
<td>388</td>
<td>1,619</td>
</tr>
<tr>
<td>Total</td>
<td>1,003</td>
<td>1,023</td>
<td>1,019</td>
<td>3,045</td>
</tr>
</tbody>
</table>

Table 2-10 – Breakdown of SP Interviews by Mode and Survey Approach (Burge, Kim, & Rohr, 2011)

Two stated preference surveys were asked for each respondent, the first related to choices between currently available options (car, air, and rail) and the second added hypothetical high speed rail service as a fourth option. For each alternative, there were a number of attributes including:

• Journey time, which included separate components for local access, local egress, station wait time, and in vehicle travel time for rail and air based trips.
• Journey time variability, which is defined as the percentage of trips that arrive within 10 minutes of the expected arrival time.
• Rail and air frequency, which is the expected headway for consecutive departures.
• Travel cost, which were displayed as single or return fares depending on the respondent. Standard and first class costs were separated. If a mode did have two fare classes, each fare class is an independent choice option for the respondent.

In addition to the mode choice options, the stated preference survey also includes a “do not travel” option in the situation where none of the available options are satisfactory to the respondent. A sample stated preference scenario is listed in Figure 2-3 below.

Prior to model estimation, a preliminary mode switching analysis was done to generally assess whether or not respondents desired to change travel modes from the mode they have recently or were currently utilizing. This preliminary analysis revealed that there is a higher propensity for travelers to stay with their existing mode choice when high speed rail was not an option. When the high speed rail alternative was present, there was more trading, especially for existing rail users.

Table 2-11 below summarizes the attributes that were used in the resulting choice model. The models were set up to reflect individual choices and the costs reflect per person costs to maintain model consistency and to have correlation with the reveal preference data. If a respondent chose
the “do not travel” option, that scenario was dropped from the model. However this filtering procedure only reduced the total data set by approximately 1%.

<table>
<thead>
<tr>
<th></th>
<th>Car</th>
<th>Air</th>
<th>Rail</th>
<th>HSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to get to train station or airport</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Waiting time at airport</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time spent in car, train, or airplane</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Time to get from train station to airport</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Percentage of trips “on time”</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Service frequency</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interchanges</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowding (rail had separate crowding by class)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Total travel cost (standard class)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Total travel cost (first class)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2-11 – Summary of Examined Attributes in SP Choice Experiment (Burge, Kim, & Rohr, 2011)

Some initial assumptions in model estimation was the simplified assumption that the observation within the data set are independent despite the fact that each respondent answered multiple SP scenarios consecutively; however, this assumption shortened modelling run times. Final models corrected for this assumption by applying bootstrap re-sampling procedures to obtain correct error estimates.

The resulting model suggests that the utility of high speed rail differs significantly depending on the mode of travel that the respondent was originally using for his or her journey. When originally traveling by rail, observed mode-switching behavior in the SP survey is a result of differences in level of service. When originally traveling by car and air, there were positive and significant constants on high speed rail; however, the extent of significance was not clear.
Figure 2-4 above reveals the nesting structure used in the final model estimations from both stated preference surveys. Conclusions from this study found that there are higher cross-elasticities between rail and high speed rail than between public transport modes and automobile.

2.4 Web-based Survey

Traditional survey administration modes have been conducted through mail, intercept, and telephone alternatives. For example, the 2006 Transportation Tomorrow Survey (TTS) has been conducted primarily through the mailing of an advanced letter stating the purpose of the survey followed by a telephone based interview (Data Management Group, 2010). However, the increase in telemarketers has reduced the effectiveness of legitimate telephone surveys. This reduction can be seen in the decreasing response rate over TTS programs through the years.

With the increasing adaptation of computers with access to the internet, email or web-based survey methods have emerged as another alternative to collect preference information. Currently, the 2011 TTS has incorporated a web-based option for households to complete a survey if they do not wish to be contacted via telephone. Given the multitude of survey administration modes available, to date, no single mode has been proven to be unambiguously superior to the others (Fleming, 2009).
The main benefit of web-based surveying is the lower cost of data collection compared to traditional mail and telephone alternatives. The comparatively low cost of web-based surveys enables larger sample sizes to decrease sample variance (Fleming, 2009). Primary areas of cost savings are associated with the reduction of printing and mailing needs. In addition the time requirement for coding, entering and verifying data is reduced when using web-based data collection (Amar, 2008). Electronic survey design also allows for the inclusion of media (sound, images, video) and interactive elements (skip logics, filtering) to make the survey experience as unobtrusive as possible (Amar, 2008).

However, there are challenges associated with web-based surveys. Exclusive use of web-based surveys might exclude individuals aged 65 or older, who have a higher chance of having difficulties with technologies, inducing sample bias. In addition, the nature of web-based advertisements has created a cultural trend where emails from unknown addresses are often viewed as spam and ignored (Amar, 2008). Another disadvantage is a non-response bias that is introduced when respondents within the sample frame have different attributes and/or demographic characteristics to those who do not respond, such as when different levels of technical ability are present among potential respondents (Fleming, 2009).

A conventional mail and web-based survey was conducted to obtain travel cost information from visitors to Fraser Island in Australia. Mail surveys were distributed in pre-paid self-addressed envelopes while the web-based survey was distributed from the School of Economics website as well as Fraser Island tourism information websites (Fleming, 2009). The mail survey yielded a gross response rate of 34.0% with 7.1% of completed surveys being invalidated. The web-based survey yielded a gross response rate of 39.1% with 15.6% of completed surveys being invalidated. In addition, the socio-demographic profiles from both survey modes were not significantly different (Fleming, 2009). From this case study, it was found that web-based survey methods are a viable alternative for conventional mail surveys at a lower cost.

Another case study to assess the effectiveness of web-based surveys was a questionnaire distributed to students and faculty of Michigan State University (MSU) about a watershed development plan. This survey was both hard copy mailed and emailed to separate samples of students within MSU (Kaplowitz, Hadlock, & Levine, 2006). The results of the case study found that the mail-based survey saw the best response rate (31.5%) but at a cost of $10.97 per
response. However, a postcard and email-based survey still achieved a high response rate (29.7%) at a much lower cost of $1.31 per response. An email only survey had the lowest response rate (20.7%) for $1.32 per response (Kaplowitz, Hadlock, & Levine, 2006). Regarding the statistical properties of the responses for each different mode, there were statistically significant differences in mean response to variables; however, they were small in absolute size (Kaplowitz, Hadlock, & Levine, 2006). It is important to note that in this case study; both the mailing list and the email list were readily available through university databases.

There are no established methods to develop web-based surveys (Alvarez & VanBeselaere, 2005); however, two basic approaches (probability and non-probability surveys) have been developed by Mick P. Couper from the University of Michigan. The non-probability approach makes no attempt to identify the sample frame or to randomly select respondents. This approach is typically used when it is not necessary to identify members of a target population or when it is difficult to contact a probabilistic sample from the population (Alvarez & VanBeselaere, 2005). Examples of non-probability surveys would include web surveys for entertainment purposes on websites such as gotoquiz.com, quizrocket.com, or even “question of the day” polls from major news media portals. While these polls offer some sort of tallied response, they do not lead to generalizations beyond reflecting the views of respondents and are typically intended for entertainment purposes (Couper, 2000).

A second type of non-probability web survey use open invitations on popular websites or internet portals. These surveys typically do not have access restrictions and have little control over multiple completions.

### 2.5 Web Recruitment

The claim to have recruited a representative sample population is often the basis of argument regarding the validity of research findings (Thomas, Bloor, & Frankland, 2007).

During the survey design procedure, a sample frame is defined to best match the population of the study area. A convergence error occurs when there is deviation between the sample frame and the target population (Alvarez & VanBeselaere, 2005). If the degree of deviation between the sample frame and population is too high, then the statistical relevance of the collected information is questioned. For example, if a survey was designed to collect information
regarding a specific university course, the expected convergence error is minimal as university email addresses or in-class interviews could be used as the sample frame. However, implementing a web survey of a large group, such as trip makers in the Greater Toronto and Hamilton Area, would result in significant convergence error because there is no available list of emails and not every trip maker would have internet access. Convergence error can be reduced by understanding the source of error as well as development of sample recruitment techniques that aims to capture a statistically significant profile of the population.

Currently, one source of convergence error is internet penetration. Internet penetration is the accessibility of the internet by computers, mobile phones, or other communication tools within a given area. The 2010 Canadian Internet Use Survey concluded that 79% of Canadian households had access to the internet (Statistics Canada, 2011). For households with internet access, most used a desktop computer (71%) or laptop (64%) followed by wireless handheld device (35%) and game console (20%) (Statistics Canada, 2011). In addition to internet availability, Canadians also have been increasing the duration of internet access. On average, an average Canadian user spent 25 minutes per day on the internet in 2000, increasing to 84 minutes per day in 2007 (Industry Canada, 2011).

Of the households that did not have internet access (21% of Canadian households), respondents stated no interest (56%), lack of device to access internet (15%), and lack of confidence, knowledge, or skill (12%) (Statistics Canada, 2011). In addition, households within the lowest income quartile stated the cost of service or equipment as a reason to not having internet access (24%) (Statistics Canada, 2011). While there has been visible growth Canadian internet penetration, it is also important to consider how individuals without household internet access might impact the convergence error of a web-based survey.

From the above statistics on internet penetration, there are two problems to web-based surveys. The first problem occurs because all potential respondents within the target population are not within the frame population as households without internet access might be omitted. The second problem is the necessity to construct a suitable frame that is statistically representative of the population with the limitations imposed during web-based surveys (Couper, 2000). Unlike household telephone numbers, which is a ten-digit numerical combination with the first three digits designating a general geographic area, it is nearly impossible to generate a list of email
addresses of a target population without sending unsolicited emails beyond the required population scope (Alvarez & VanBeselaere, 2005). Other issues that researchers have encountered with web-based surveys have been self-selection bias, sampling bias, costs of random sampling and technical problems such as computer incompatibilities (McMellon & Schiffman, 2011).

In addition to potential convergence error associated with the sample selection, there is another type of error associated with sample recruitment when no information is collected from a sampled unity due to refusal or non-contact called unit non-response (Barriball & While, 1999). Previous research into unit non-response has suggested that the importance and usefulness of a proposed survey should be conveyed onto the respondent to influence the respondent’s decision to participate (Barriball & While, 1999). Face-to-face contact with potential respondents is another important aspect to reduce unit non-response (Barriball & While, 1999); however, face-to-face interaction would be limited in a large population and the use of web-based medium does not facilitate the use of personalized contact.

Burnfield from Bowling Green State University designed and completed a survey to examine meeting characteristics and the perceptions of meetings. For this study, a number of recruitment methods were utilized including; a research participant panel from StudyResponse, snowball sampling, university alumni newsletters, and email distribution lists (Burnfield & Rogelberg, 2003). A multitude of recruitment techniques were used to maximize the likelihood of obtaining a diverse international sample that would potentially represent a range of occupations. For the StudyResponse panel, a raffle incentive of $50 was offered while respondents from other recruitment methods were not offered any incentive (Burnfield & Rogelberg, 2003).

Emails were sent to 3400 potential respondents in the StudyResponse panel in ten (10) waves. If the respondent did not complete the survey, follow-up email reminders were sent one (1) week after the initial wave and two (2) to three (3) weeks after the final wave. Approximately 2700 reminder emails were sent out during the course of Burnfield’s survey. For the first survey, a sample size of 468 was achieved after removing ineligible participants. From the total 468 respondents, 378 (80.8%) was from StudyResponse and 90 (19.2%) was from other recruitment sources (Burnfield & Rogelberg, 2003). For the second survey, a sample size of 1,209 was achieved after removing ineligible participants. From the total 1,209 respondents, 431 (35.6%)
was from StudyResponse and 778 (64.6\%) was from other recruitment sources (Burnfield & Rogelberg, 2003). StudyResponse respondents were identified by a unique ID string that was verified upon the conclusion of the web-survey.

The different allocation of StudyResponse panel respondents and other respondents could have been attributed to the location of when the StudyResponse ID was requested. In the first survey (80.8\% StudyResponse panel), an ID number was requested at the beginning of the survey in comparison to the second survey (35.6\% StudyResponse panel) where the ID was requested in the last question of the survey (Burnfield & Rogelberg, 2003). In addition, it was recommended that close-ended and open-ended responses should be employed within survey questions to determine the recruitment strategy associated with a particular respondent in addition to requesting information such as a ID number. This use of closed and open ended response would help the survey designer and subsequent researchers understand sample composition as well as if respondents were recruited from more than one recruitment source (Burnfield & Rogelberg, 2003).

The increased accessibility to the internet has resulted in the development and wide-spread adaptation of web-based social networks. Currently, the largest of these web-based social networks have been Facebook, Linkedin, and Twitter. In addition to traditional forms of advertising, companies have started creating relationships through the use of different social networks as another form of advertisement. Social networking strategies by these large and small companies are typically used to either advertise a product or to recruit potential employees; however, parallels could be made between the recruitment of employees and the recruitment of respondents for web-based surveys.

A case study of active internet users over the age of 55 was completed to assess practical methods of sample recruitment and database development of a web-based survey where a small number of respondents were drawn from a larger general population (McMellon & Schiffman, 2011). The objective of the case study survey was to examine the allocation of internet activities as well as personal characteristics of internet use. A web-based recruitment process was utilized followed by a mail-in questionnaire. At the time of the case study (2001), internet use was not as widespread compared to the present and literature on internet-based recruitment and research was not readily available. The basis of McMellon’s research was primarily based on a project on
internet chat-group participation (McMellon & Schiffman, 2011). The case study was broken down into two stages; a recruitment stage followed by an administrative stage.

The first stage attempted to recruit from a population of adults, over the age of 55, who were active on the internet. The difficulty encountered during this stage was attempting to draw a random sample frame from the population without a developed mechanism for web-based random sampling (McMellon & Schiffman, 2011). The available alternatives were to recruit respondent using mail-based recruitment from the larger population (regardless of age or internet use) or using user lists from online services and local access providers. The first alternative was not chosen as the target population constituted about 6% of the online community and a mass mail-based recruitment process would be too costly to implement. The second alternative was not chosen when internet service providers declined to provide McMellon with a subscriber list (McMellon & Schiffman, 2011). With the difficulties developing a cost-effective method of sample recruitment, McMellon recruited respondents directly from the internet.

A multistage recruitment process was also utilized given recommendation of previous research which suggests multistage recruitment processes when it is difficult to create a list of the total population (Fowler, 1988). The mailing list from a number of online groups were obtained and aggregated into a single e-mail list. Three lists were utilized for the case study; 3,487 from SeniorNet, 979 from CompuServe’s American Association of Retired Persons (AARP), and 455 from CompuServe’s retirement issues forum (McMellon & Schiffman, 2011). From the total 4,921 emails sent, there were 1,003 undelivered emails due to incorrect address copied, users dropping the service, full mailbox, and change of online address. A second email was sent when individuals did not respond to the original email; however, this second email was met with a number of complaints (McMellon & Schiffman, 2011). Due to a negative response by the population upon receiving a second recruitment email, the researchers abandoned further use of reminder emails and letters; reducing any potential increase in response rates at the conclusion of the survey.

From the recruitment process, a total of 369 individuals responded positively to the recruitment email and 284 questionnaires were completed and returned to the researchers (McMellon & Schiffman, 2011). Compared to traditional survey with a response rate ranging from 20-25%, the 5.7% response rate from this case study is considered to be low. The low response rate from the
case study might have resulted from; unfamiliarity with internet technology, respondent ignoring email, email was never delivered, respondent did not want to respond, and a level of paranoia regarding confidentiality (McMellon & Schiffman, 2011). McMellon has aggregated a number of different studies as a benchmark to the case study seen in Table X below. For studies that compared regular mail and web-based surveys, the response rate should be approached with caution as the studies do not take into account cost, speed, subject matter or target population, which could influence factors beyond raw response rates (McMellon & Schiffman, 2011).

A non-probabilistic sample was obtained using the case study procedures. There were initial concerns that the collected database was not representative of the three online groups. Manova analysis was performed between the groups. The Manova analysis revealed that the combination of the three groups had acceptable aggregation (McMellon & Schiffman, 2011). While there was acceptable aggregation of the database, the use of only three online groups might not represent the target population of active internet users over 55 years old. In addition respondents who responded positively to the recruitment email might be having a higher level of risk-taking than those who did not respond (McMellon & Schiffman, 2011).

Thomas from the University of Sussex drew from four different research studies to address the process of sample recruitment as well as the gaps between methodology and implementation of sample recruitment. The four research cases were a combination of qualitative and quantitative non-probabilistic studies using varying direct and indirect sample recruitment techniques (Thomas, Bloor, & Frankland, 2007). An abundance of research has been completed to address the biases associated with response error, little research have been completed to address the social processes of sample recruitment (Thomas, Bloor, & Frankland, 2007). An earlier study on gathering sociological data concluded that there were no universal rules of sample recruitment that could be used for all occasions (Gephart, 1988) despite the continued use of recruitment rules and the assumptions that these rules are followed during actual sample recruitment practice.

A study of the sexual health of cruise ship crews recruited male and female crew members in two major cruise ports. Upon eligibility, a questionnaire was administered to consenting respondents immediately after recruitment (Thomas, Bloor, & Frankland, 2007). During this study, researchers recruited respondents based on recognized cues in appearance and social context. It was quickly recognized that a completed survey was more successful if approaching lone crew
members rather than couples or parts of a group (Thomas, Bloor, & Frankland, 2007). The expectation of rejecting makes approaching individuals more difficult. In addition, researchers documented some hesitation in sample recruitment when there was a potential to recruit an individual more than once. As a result, it was noted that researchers were more inclined to approach individuals with distinguishable features or characteristics because those individuals would be more memorable.

The public’s use of community pharmacy sampled women over the age of 65 and mothers of pre-school aged children who did not have full time employment. The major recruitment source was cold-contacting potential respondents along shopping streets in three contrasting socio-demographic areas (Frankland, 2002). A short time was available for the researcher to assess the eligibility of a potential respondent; however, Frankland was concerned about offending individuals who might not fall under the old age eligibility.

A study on male prostitution in Glasgow recruited 28 street-working male prostitutes primarily through snowballing. Efforts were made to reduce cold-contacting as a recruitment strategy because due to the potential for misidentification (Bloor, Barnard, Finlay, & McKeeganey, N., 1993). The initial groups of male prostitutes were recruited by observing men who lingered in and around lavatories; however, long observation periods were required (240 hours of total fieldwork) to potentially identify valid respondents (Thomas, Bloor, & Frankland, 2007). Especially with sensitive research subject such as prostitution, there is a change where the respondent might lie about their eligibility. For example, some men approached in this study claimed they were ineligible when cold-contacted, only to be recruited again when snowballing from another contact (Bloor, Barnard, Finlay, & McKeeganey, N., 1993).

The fourth study in Thomas’ paper was about a woman’s experience of new sexual relationships abroad. Women were initially asked to be a part of the study at the departure lounge of a regional airport and a telephone call upon arrival confirmed eligibility for the study (Thomas, Bloor, & Frankland, 2007). Thomas addressed concern about people taking offence when asked about their private life and this was reflected on the potential respondents initially approached at the departure lounge. In this study, there was a probable bias as the researcher felt more inclined to approach groups of young women for the fear of causing offence in older women as their relationship status was more ambiguous (Thomas, Bloor, & Frankland, 2007).
2.6 Literature Review Summary

Compared to the availability of data and models on intracity travel demand, there is currently a gap in the transportation research field regarding intercity travel demand. One reason is the issue regarding political jurisdictions as intercity travel information requires cooperation of municipal, provincial and even federal agencies.

The use of web-based survey instruments is a relatively new frontier of research-based data collection. The recent adaptation of web-based surveys by data collection projects such as the Canadian Census and the most recent TTS program is a result of the increased spread of internet availability. However, in most cases, a web-based survey has only been a supplement to the existing survey program. For example, the 2011 TTS will have a web-based survey component, but phone lists and advance letters were still required to inform the interviewee beforehand. In addition, the email only survey scenario at MSU had available email listings from university administration (Kaplowitz, Hadlock, & Levine, 2006). For web-based surveys, the current gap is the combined effectiveness of new survey recruitment techniques in the collection of data via web-based instruments.
Chapter 3
Survey Design

The proposed survey design aims to collect a wide variety of information regarding a potential respondent’s local and intercity travel behaviors as well as relevant socioeconomic attributes. As a large number of questions may be required of a respondent, the survey flow is important to reduce the possibility of survey fatigue. As a result, the proposed survey design can be broken down into four (4) separate sections; intercity revealed preference, intercity stated preference, local daily revealed preference, and socioeconomic questions. The full set of questions potentially available to the respondent can be found in Appendix C. The following sections summarize the purpose and goals of each individual survey subset.

3.1 Intercity Revealed Preference

Where many existing travel surveys inquire about a respondent’s daily local travel/activity information, it was necessary to also inquire about a respondent’s previously intercity travel information given the project scope and objectives. Aside from matching revealed intercity travel data with existing data sets, the collection of this data also allows us potentially model if a respondent lived trip tendencies would have an effect on their stated preference choices.

The first thing asked of the respondents is their intercity trip frequencies. From existing Canadian data sets such as 2004 Canadian Travel Survey (ITS) and 2010 Travel Survey of Residents of Canada (TSRC), the frequencies of intercity travel at different distances (regional, provincial, national) have not been inquired. Collection of non-local trip frequencies may be used as a possible attribute that could affect modal choice. If an individual typically travels nationally by air, there may be a higher tendency for that individual to also travel by air for intercity destinations. Additionally, this information may useful as a baseline tool for future intercity travel surveys and research.

Next, the respondent is inquired about their most recent trip within the Quebec City/Windsor corridor within the past 12 months. As the scope of this project is within this geopolitical area, detailed information of trips to other destinations would not be of the best use in modelling efforts and may induce survey fatigue early in the survey flow. The 12 month temporal limit is
included to reduce any possible uncertainty that may be induced when inquiring about trip details beyond the time limit.

If a respondent has made an intercity trip within this corridor within the past year, basic travel information is collected including the destination city, travel mode, travel cost, and trip purpose. Similar to the intercity travel frequency questions, these responses can either be used as attributes in the intercity mode choice model or baseline data to compare to existing and future data sets also related to intercity travel. As a web-based survey instrument is used, it was possible to add in conditional questions such as vehicle rental costs or access/egress modes if the respondent answered in such a way to activate these conditional logic steps. If a respondent has indicated that no intercity trips were made within the past 12 months, then the respondent is not shown and do not have to respond to the intercity revealed preference set of questions.

3.2 Stated Preference

3.2.1 Background

The section most useful to modelling mode choice is the intercity mode choice stated preference section. While revealed preference information indicates current travel patterns, the potential modal shift due to the introduction of a currently non-existent high speed rail mode cannot be modeled using RP data. As previously stated, the objective is to obtain relevant data to model the elasticities of existing and hypothetical mode attributes. The basis of the stated preference survey is presenting the respondent with a hypothetical intercity travel scenario and inquiring about the respondent’s mode choice, out of five possible alternatives, in each of the six different scenarios. The following sections detail the different aspects related to designing the stated preference questions.

3.2.2 Hypothetical Situation

A hypothetical intercity trip between an individual’s specified home location in the Greater Toronto Area (GTA) to a location in Montreal is given. Montreal was chosen as the destination location because it is estimated that the greatest percentage of intercity trips originating from the GTA, within the Quebec City/Windsor corridor, has a destination in Montreal. Unlike existing travel surveys that record origins at the cities level, a higher degree of origin and destination disaggregation is required to provide specific access and egress attributes. The origin location is
aggregated to the Forward Sortation Area (FSA), which are the first three digits of a postal code. Similarly, the destination location is aggregated to the 19 boroughs in Montreal. A higher degree of aggregation was chosen for the destination as it was hypothesized that trip makers would not be as knowledgeable about his or her destination FSA address.

Further detailed in section 3.2.6.1, the disaggregation of local origins and destinations resulted in the necessary creation of tables for each origin-destination pair. Given a limited time to design the survey, the resources to create the multitudes of OD pairs were not available in the scope of this project. As a result, the GTA to Montreal pair was selected based on the hypothesis that the highest probability of respondents would have traveled to Montreal compared to other cities within the Quebec City – Windsor corridor.

### 3.2.3 Procedure

Given the hypothetical situation, the interviewer is prompted to make a choice from the alternatives. The alternatives are established by the survey designer and associated with variables related to the alternatives and scope of the survey. For this survey design, six (6) alternatives are given to the interviewer. A detailed description of each alternative is listed below.

### 3.2.4 Alternatives

There are a total of five mode choices available to the respondent for each of the six stated preference scenarios. Of the five mode choices, the high speed rail mode is hypothetical; however, the attributes related to this mode are sourced from the latest feasibility report from EcoTrain (EcoTrain, 2011). The following Table 3-1 indicates the mode, any possible companies providing the transit, and relevant trip information. It is assumed that the trip is a direct trip with as few layovers as possible. In addition, costs for each alternative were determined by online pricing from respective companies for the same day of travel on Wednesday, December 14, 2011.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Car</th>
<th>Bus</th>
<th>Rail</th>
<th>Air</th>
<th>HSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Company</td>
<td>-</td>
<td>Greyhound</td>
<td>VIA Rail</td>
<td>Air Canada Westjet</td>
<td>-</td>
</tr>
<tr>
<td>Departures Per Day</td>
<td>-</td>
<td>10</td>
<td>8</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Baseline Travel Time</td>
<td>6h 20m</td>
<td>7h 27m</td>
<td>5h 50m</td>
<td>1h 10m</td>
<td>3h 00m</td>
</tr>
</tbody>
</table>
The following section details each of the five (5) intercity travel mode alternatives separately including some level-of-service details as well as initially predicted results.

### 3.2.4.1 Automobile

Automobile defined as a personal vehicle with a seat capacity of five (5) persons. The automobile only requires stops when refueling or when the driver needs to take a break. If the automobile alternative is chosen, the entire trip will be made in the automobile. An automobile trip originates from the respondent’s home location and the destination is the chosen borough location in Montreal. There is no access or egress times or costs associated with the automobile mode. The travel time from the origin to the destination is influenced by the amount of traffic on the various roads and does not account for any predicaments such as adverse weather conditions. The baseline travel cost is estimated as a function of travel distance to expected fuel costs per kilometer. Expected fuel costs per kilometer was calculated by taking the average of three different types of vehicles using a national average gas price of 1.23 cents per liter (Canadian Automobile Association, 2012). Table 3-2 below is an excerpt from Canadian Automobile Association’s 2012 driving cost report.

<table>
<thead>
<tr>
<th></th>
<th>Civil LX</th>
<th>Camry LE</th>
<th>Equinox LT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel</td>
<td>8.88 cents</td>
<td>10.12 cents</td>
<td>11.36 cents</td>
</tr>
<tr>
<td>Maintenance</td>
<td>2.18 cents</td>
<td>4.60 cents</td>
<td>2.55 cents</td>
</tr>
<tr>
<td>Tires</td>
<td>1.84 cents</td>
<td>1.96 cents</td>
<td>2.53 cents</td>
</tr>
<tr>
<td>Total</td>
<td>12.90 cents</td>
<td>16.68 cents</td>
<td>16.44 cents</td>
</tr>
</tbody>
</table>

Table 3-2 – Average Annual Operating Costs per Kilometer - Variable

The automobile specifications are listed below:
- CMA to CMA Distance: 542 km
- CMA to CMA Travel Time: 6 hours 18 minutes
- Travel Costs: $56.00

This alternative is included as the majority of passenger trips within the Quebec-Windsor Corridor are automobile based (IBI Group, 2002). One of the study objectives is to determine if level of service improvements from other alternatives may shift trip makers away from the
automobile mode. The automobile mode is predicted to encompass the largest modal share, especially for households with auto ownership. The automobile modal share is expected to decrease for households who do not have a vehicle at home.

### 3.2.4.2 Bus (Greyhound/Megabus)

The second alternative used in the SP survey is to travel by bus. Currently travel by bus between the GTA and Montreal is provided by either Greyhound or Megabus. Both providers offer similar ticket price and travel similar routes from Toronto to Montreal. Depending on the time of travel, the bus alternative will make multiple stops to service local stations along the route. Currently, the access time is determined as the travel time (on the specified access mode) from the respondent’s home location to the bus terminal located on Dundas and Bay. Similarly, egress time is determined as the travel time (on the specified egress mode) from the downtown bus terminal in Montreal at 1717 Rue Berri Street to the chosen borough centroid location.

It is understood that there may be intermediate Greyhound or Megabus access stations within the Greater Toronto Area. These were not included as potential access stations due to the added complexity with understanding the catchment area that each smaller access station may impose. The bus specifications are listed below:

- CMA to CMA Distance: ~600 km
- CMA to CMA Travel Time: 6.5 (Megabus) to 8 (Greyhound) hours
- Travel Cost: $29 (Megabus) to $31 (Greyhound)

Given the availability of bus transportation between cities within the QWC, the bus alternative was chosen given its commercial viability with transporting passengers. The comparatively low price of travel for the bus mode is one of the benefits despite the longer travel time. A younger aged demographic (18-24 years) is expected to have a high bus modal share due to smaller percentage of auto ownership.

### 3.2.4.3 Rail (Via Rail)

Via Rail provides one of the last nationwide passenger rail alternatives available in Canada. Being on a dedicated right of way, the rail alternative has the potential to have the fastest station-to-station travel times compared to other land-based transit travel alternatives that share the
highway right of way. Currently, the access time is determined as the travel time (on the specified access mode) from the respondent’s home location to Union Station in downtown Toronto. Similarly, egress time is determined as the travel time from Central Station in downtown Montreal to the chosen borough centroid location using the specified travel mode. The Via Rail specifications are listed below:

- CMA to CMA Distance: ~600 km
- CMA to CMA Travel Time: 5.5 (low) to 7.5 (high) hours
- Travel Cost: $50 (low) to $80 (high)

The rail mode is included as it is a closer comparison with the hypothetical high speed rail travel mode. The level-of-service attributes of the rail alternative can be used as a potential baseline to the hypothetical high speed rail alternative. In addition, intercity travel via the rail alternative is a currently available choice for trip makers with comparative prices and travel times to bus and car alternatives.

3.2.4.4 Airplane (Westjet/Air Canada)

The fourth alternative of travel is by airplane. The two airlines listed above have similar pricing and travel time. For the scope of this project, air travel is made from Pearson Airport with services provided by air Canada and Westjet. Flights departing from Billy Bishop Airport are not included because of the added complexity of defining the catchment area of both airports or providing two air travel based modes.

Compared to the other modes previously described, the airplane alternative would have the fastest travel time. Given the faster travel time, the access time (getting to the airport) and egress time (traveling from the airport to the destination) may have an effect on modal choice as it influences the total travel time when considering the population distribution. An additional level-of-service attribute relatively unique to the airplane alternative is the additional time required to check in and clear security when departing from the origin city as well as the additional time required picking up luggage when arriving at the destination city.

The airplane trip specifications are listed below:

- Distance: ~525km
- Travel Time: 1 hour 15 minutes
- Travel Costs: $180 (low) to $280 (high)

The airplane mode was chosen to potentially exhaust the major travel alternatives given the above-listed hypothetical situation posed for the stated preference survey. Compared to the previous travel mode alternatives, the airplane mode has the highest cost and the lowest travel time, it is expected that the automobile modal share is dependent on the purpose of the trip. It is expected that business-based trips would have a larger airplane modal share compared to personal or social travel purposes.

3.2.4.5 High Speed Rail

The fifth and final alternative presented to the respondent for the SP survey is the hypothetical high speed rail mode. The travel time, location of station, and other relevant attributes related to this mode were taken from EcoTrain’s feasibility report. A summary of the EcoTrain report is detailed in section 2.3.1. The EcoTrain report was used as it was the most recent and comprehensive study on this potential travel mode.

<table>
<thead>
<tr>
<th>City Pair</th>
<th>Existing VIA Rail Service</th>
<th>F200+</th>
<th>E300+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quebec City – Montreal</td>
<td>3h 09m</td>
<td>1h 49m</td>
<td>1h 26m</td>
</tr>
<tr>
<td>Montreal – Ottawa</td>
<td>1h 55m</td>
<td>1h 11m</td>
<td>0h 57m</td>
</tr>
<tr>
<td>Ottawa – Toronto</td>
<td>4h 36m</td>
<td>2h 25m</td>
<td>1h 50m</td>
</tr>
<tr>
<td>Montreal – Toronto</td>
<td>5h 12m</td>
<td>3h 38m</td>
<td>2h 47m</td>
</tr>
<tr>
<td>Toronto – London</td>
<td>2h 07m</td>
<td>1h 05m</td>
<td>0h 51m</td>
</tr>
<tr>
<td>Toronto - Windsor</td>
<td>3h 59m</td>
<td>2h 12m</td>
<td>1h 33m</td>
</tr>
</tbody>
</table>

Table 3.3 – Travel Times Between Major City Pairs

<table>
<thead>
<tr>
<th>City Pair</th>
<th>Existing VIA Rail Service</th>
<th>F200+ (2025 projection)</th>
<th>E300+ (2025 projection)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quebec City – Montreal</td>
<td>4</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>Montreal – Ottawa</td>
<td>6</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Ottawa – Toronto</td>
<td>5</td>
<td>22</td>
<td>25</td>
</tr>
<tr>
<td>Montreal – Toronto</td>
<td>7</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Toronto – London</td>
<td>5</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Toronto - Windsor</td>
<td>4</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

1 Taxes, fees, charges, and surcharges included
For the stated preference survey, the departure and arrival stations are similar to the rail travel mode. The high speed rail departs from Union Station in downtown Toronto and arrives at Central Station in downtown Montreal. While the EcoTrain feasibility report indicated that there may be intermediate stops even within the Toronto and Montreal central metropolitan area, the downtown core departure and arrival stations were chosen to reduce the complexity of considering each station’s individual catchment area.

The high speed rail specifications are listed below:

- Distance: ~550 km
- Travel Time: 2 hours 45 minutes (anticipated)
- Travel Costs: 1.8 x VIA Rail fares (anticipated)

As high speed rail service does not currently exist in Canada, this is a hypothetical mode choice that is presented to the respondents. The baseline values were selected as such to best represent a potential high speed rail trip. When priced competitively between the airplane and traditional rail travel costs, some mode shift to high speed rail is expected. It should be noted that any mode shift to high speed rail may be attributed to a respondent's interest in trying a new travel mode and may not represent actual ridership values. This bias in new technologies is inherent in this type of survey; however, the data collected should still be assessed as valid.

### 3.2.5 Attribute and Levels

In the SP survey, each situation presented to the interviewer will be a table (or text representative) of alternatives with corresponding attributes for each alternative. At the pilot stage, an exhaustive list of alternatives was utilized. For the final design of the survey, some attributes were removed from the stated preference design. The full exhaustive list is detailed below with indication to whether or not that attribute was used for the final survey design. For the construction of this intercity survey, three questions were asked when deciding the attributes required:

1. Why is this attribute chosen?
2. What are the levels associated with the attribute?
3. What assumptions were made?
4. What is the perceived contribution in the results?
5. Is the attribute generic or alternative specific?

The first question is an initial validation for why the attribute was chosen and helps to establish reason for why the attribute should be included. Perceived contribution is an estimation of the importance (in term of high, medium, low) of an attribute to the resulting model from the collected data. This step is important as an estimation tool for the researcher and used as a tool to generate more attributes. The final question, regarding generic or alternative specific attributes, is related to the construction of the utility functions and will be described in section 3.2.8.

After each attribute was selected, levels were assigned to each attribute. Levels are the different values that can be associated for each attribute for a given alternative. Levels can either be relative values associated with the attribute (such as a dollar amount with a travel cost attribute) or a generic value that is given a description (such as 0 representing crowded and 1 representing not crowded). In the resulting stated preference design, attributes may vary in each of the six scenarios. By varying the values of some attributes, it may be possible to see which of the selected attributes can be mathematically tied with a respondent’s modal choice. If an attribute is not statistically significant to modal choice, then varying that specific attribute would not result in the respondent shifting modal choices due to a change in that level-of-service attribute.

### 3.2.6 Selected Attributes and Levels

A set of attributes was generated to provide the respondent relevant information regarding each travel mode. An initial list of attributes was based on a TSRB study on a high speed rail line in London, UK (Burge, Kim, & Rohr, 2011) and modified to better relate to the Canadian region and existing Canadian travel mode alternatives. Different levels are the different values that can be associated for each attribute. Levels can either be relative values associated with the attribute (such as a dollar amount with travel cost) or a generic value that is given a description (such as 1 booked seating and 2 representing arranged seating). Table 3-5 below is a list of the attributes used in the proposed survey design with the alternatives associated with the specific attributes and the level used in designing the stated preference survey.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Available Modes</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access Mode</td>
<td>B/R/A/HSR</td>
<td>Car/Transit/Taxi/Non-motorized</td>
</tr>
</tbody>
</table>
With the list of attributes and its respective levels established, an orthogonal stated preference design was completed using Ngene software. The basis of using Ngene was to compute a subset of possible SP scenarios that would yield the most relevant results. The Ngene software is described in great detail in section 3.2.8.

It is understood that only main effects are considered with an orthogonal design versus the main and interaction effects of efficient stated preference survey design; however, there was a lack of existing models that incorporated the proposed attributes as baseline coefficients for the initial stated preference design.

### 3.2.6.1 Access Mode

The access mode attribute is defined as the local mode choice option that the respondent would use to travel to the departure station of an intercity travel mode alternative. This access trip should originate from the respondent’s home location (aggregated to the geographical center of the respondent’s home FSA address) and end at the location of the intercity mode’s departure station. In conjunction with access time, the inclusion of the access mode, as a potential attribute of intercity mode choice, is to assess whether or not the local accessibility of transportation alternative has an effect on a respondent’s intercity mode choice. The intercity mode destination stations in the GTA are:

- Greyhound and Megabus busses departs at Toronto Coach Terminal at Bay & Dundas
- Via Rail and high speed rail trains departs at Union Station at Front & Bay
• Air Canada and Westjet airplanes departs at Toronto Pearson Airport at 6301 Silver Dart Drive

The possible access mode alternatives are typical travel methods currently used to access the designated intercity departure stations; however, the list is more generalized to keep the number of alternatives under four possible levels. This generalization was done due to the current limitation of having only six distinct scenarios. If more access mode alternatives were presented, then an increased number of stated preference designs would be required to find statistical significance in the survey results.

The cost of utilizing various access modes is also presented to the respondent and is a stratified price based on the access mode. The decision to include a generalized range of travel cost was done to simplify the total number of attributes presented to the respondent. In addition, it was assumed that travel cost in each access mode alternative were mutually exclusive. For example, the cost of transit would range between $3 to $6 whereas a trip by taxi with the same origin and destination would be over $10. The local access modes and corresponding prices are listed below.

• Drop-off (Cost = $0)
• Transit (Cost = $3 to $6)
• Taxi (Cost = $10 to $30)
• Non-motorized (Cost = $0 to $5)

Some assumptions made when designing the access mode attribute was that a respondent would only require one access mode to get from their home location to the intercity mode departure station. In reality a respondent may require multiple access modes, especially if his or her home location was farther away from major population centers. However, incorporating multiple access modes would introduce more complexity into the SP design and it was assumed that most respondents would relate to a main access mode. Another assumption made was the lack of park and ride option. This park and ride option was not included as this alternative is only currently available at the airport and the price varies with the number of days required.

The access mode is not expected to have a great influence on intercity modal choice. While there may be some influence, it is most likely expected to the corresponding access time, which may
influence the overall travel time for the respondent. It is hypothesized that respondents do not consider the access mode unless the price of accessing a departure station greatly increases the overall travel costs. For example, a respondent may not pay for a taxi ride (costing $30) to access the Megabus (also costing $30). The access mode attribute is considered alternative specific given the associating attribute levels are local travel mode alternatives.

3.2.6.2 Access Time

Access time is defined as the time required traveling from a respondent’s home location (aggregated to the centroid of the home location’s FSA address) to the specified intercity mode departure station, using the corresponding access mode. The purpose of the access time attribute is to assess if local accessibility has any influence on intercity modal choice. For example, it is relatively easy to travel to all three departure stations from a Toronto downtown location; however, a similar trip may take longer if travelling from an Ajax home.

The baseline access times were determined by using Google Maps to query directions between a FSA centroid location and the three intercity departure locations (Toronto Coach Terminal, Union Station, and Lester B. Pearson Airport). Google Maps was used as the service provided travel times by automobile, transit, as well as non-motorized transport. There were three levels associated with access travel time and each of these levels was a multiple of the baseline access time determined from Google Maps. This percentage based designation of access time was chosen over absolute levels to address the sensitivity of respondents to travel time. If absolute time (for example, ± 10 minutes) was used on a respondent originating from downtown Toronto, the difference in travel time to Union Station would seem much greater than the travel time to Lester B. Pearson airport. The associated access time levels are as follows:

- Existing access travel time (min) * 0.50
- Existing access travel time (min) * 1.00
- Existing access travel time (min) * 1.50

For automobile trips, Google Maps presented a real-time estimation of travel time based on perceived traffic volumes as well as an idealized travel time. The idealized automobile travel time was used as the baseline value. While automobile travel times are dependent on traffic
conditions, it was assumed that the coefficients used in the access time levels would account for any changes in automobile travel time.

For transit based trips, Google Maps has the option to select a specific departure or arrival time to output relevant transit information. For the stated preference design, a weekday trip arrival time of 13:00 was selected for all transit trips. This assumption was made to reduce the variability of transit schedules and to select a trip that was slightly out of peak-hour schedule but may still be affected by traffic volumes. Similar to the automobile access time, it was assumed that the variation of the attribute levels may account for increased or decreased transit travel times. Another aspect to consider in using Google Maps to determine transit travel times is the use of competing transit services. In the study scope of the Greater Toronto Area, a multitude of transit alternatives exist and there may be a number of travel choices that could be made in one trip. For example, a transit trip from a home location in Brampton to Union Station could be made via Brampton Transit with a transfer to TTC transit. A similar trip could also be made only using GO Transit or even a combination of Brampton Transit and GO Transit. The assumption was made that the shortest transit travel time listed on Google Maps was chosen but rounded up to the nearest 15 minute increment to account for variability in transit modal choice.

Non-motorized trips were recorded with Google Map’s suggested travel time by walking. Biking times were not available to every origin-destination pair and was assumed to be a function of the walking time as similar routes may be used. Non-motorized walking trips were recorded for every OD pair; however, many of these trip pairs had walking trip times beyond logical choices. For example, a respondent living in Pickering would have to walk for eight (8) hours to arrive at Lester B. Pearson airport. As a result, a walking time of 60 minutes was chosen as a maximum upper limit for non-motorized transport. If the stated preference design stipulated a non-motorized access mode with an access travel time exceeding 60 minutes, then the program would randomly choose from the other three access mode alternatives and output the corresponding access travel time.

The contribution of access time to intercity modal choice is expected to vary with the home location of the respondent. Given local access modes and distances to the intercity departure stations, an increased travel time may add to the total travel time of a respondent. For example, a plane trip from Toronto’s Lester B. Pearson airport may take an hour and a half; however, it may
also take two hours for a respondent living in Burlington to arrive to the airport by transit, adding an additional two hours to the total travel time. Access time would be considered as a generic attribute as access time is based on each respondent’s home location.

3.2.6.3 Egress Mode

Similar to the access mode attribute, egress mode is defined as the destination local mode choice option that the respondent would use to travel from the arrival station of an intercity mode choice alternative to a local destination in Montreal. This egress trip should originate from the location of the intercity mode’s arrival station and end at the geographic centroid of one of Montreal’s 19 boroughs. In conjunction with egress time, the inclusion of the egress mode, as a potential attribute of intercity mode choice, is to assess whether or not the local accessibility of transportation alternative at the destination location has an effect on a respondent’s intercity mode choice. The possible arrival stations in Montreal are:

- Greyhound and Megabus busses arrives at Berri Street and Boulevard de Maisonneuve East
- Via Rail and high speed rail trains arrives at Montreal Central Station
- Air Canada and Westjet airplanes arrives at Pierre Elliott Trudeau International Airport

Compared to the access alternative, there is an additional automobile rental option included within the total levels of the egress mode attribute. Similar to access mode, it was also assumed that the respondent would utilize only one egress mode when travelling between the arrival station and the destination location. If more than one egress mode is chosen by the respondent, the listed attribute is considered the primary egress mode used.

Additionally, the cost of utilizing various egress modes is also presented to the respondent and is a stratified price based on the listed egress mode. The decision to include a generalized range of travel cost was done to simplify the total number of attributes presented to the respondent. It was assumed that travel cost in each access mode alternative were mutually exclusive. The local access modes and corresponding prices are listed below.

- Pick-Up (Cost = $0)
- Transit (Cost = $3 to $6)
- Taxi (Cost = $10 to $30)
- Rental (Cost = $50)
- Non-motorized (Cost = $0 to $5)

An assumption made when designing the egress mode levels is that the costs associated with each of the egress mode alternatives is relevant to the expected minimum cost of a one-way trip and should not ideally consider for any extra travel expenses incurred during the intercity trip; such as parking. However, it may be possible for the respondent to purchase a transit day-pass or utilize the rental vehicle for a longer duration of time.

The egress mode is not expected to have a great influence on intercity modal choice. While there may be some influence, it is most likely expected to the corresponding egress time, which may influence the overall travel time for the respondent. Similar to the access mode attribute, egress mode is hypothesized that respondents do not consider the egress mode unless the price of accessing a departure station greatly increases the overall travel costs. The egress mode attribute is considered alternative specific given the associating attribute levels are local travel mode alternatives.

3.2.6.4 **Egress Time**

Egress time is defined as the time required to travel from the specified intercity mode arrival station to specified destination in Montreal (aggregated at the borough level), using the corresponding egress mode. Similar to the access time attribute, the purpose of the egress time attribute is to assess if local accessibility has any influence on intercity modal choice.

The baseline egress times were determined by using Google Maps to query directions between the three intercity arrival stations (Station Berri, Montreal Central Station, and Pierre Elliott Trudeau International Airport) and the geographical centroid of each Montreal borough. Google Maps was used as the service provided travel times by automobile, transit, as well as non-motorized transport. For automobile trips, Google Maps presented a real-time estimation of travel time based on perceived traffic volumes as well as an idealized travel time. The associated access time levels are as follows:

- Existing egress travel time (min) * 0.50
• Existing egress travel time (min) * 1.00
• Existing egress travel time (min) * 1.50

The assumptions made regarding access travel time for all local travel mode alternatives were also made for egress travel time. Similar to access travel time, non-motorized trips over 60 minutes would be replaced by a randomized choice of the other egress mode alternatives and corresponding egress travel time.

The contribution of egress time to intercity modal choice is expected to vary with the destination of the respondent. While the egress time attribute is similar to the access time attribute, egress time is not expected to have as but of an influence to intercity mode choice as access. One hypothesis is the idea that individuals are less sensitive to choosing local travel modes at the destination location compared to local travel modes at their home city. Egress time would be considered as a generic attribute as access time is based on each respondent’s destination location.

3.2.6.5 On Time

The on time attribute is defined as the probability that the respondent will arrive to Montreal on time using the chosen intercity travel mode. The purpose of this attribute is to present probable trip scenarios that account for any possible delay. In addition, another purpose is to assess whether or not the probability of delay would influence any modal shifts. For example, driving an automobile from Toronto to Montreal experience some delays depending on the volume of traffic within the highway corridor. In addition to personal vehicle, these delays may also affect the scheduled arrival time for the bus travel mode. In comparison, as rail and air operates on a dedicated right of way, there might be less probability of arriving late unless of some situation causing delay; however, there are other factors that may influence the probability of arriving to the destination on time.

It should be noted that this on time attribute corresponds with the intercity trip and does not affect the access and egress travel times. This decision was made to focus on the level-of-service attributes of the intercity travel mode rather than local travel modes.

In the proposed design, it is assumed that there is always uncertainty at the arrival status of any intercity travel alternatives. Another assumption is that the difference between the designated
levels (at 10% increments) is spaced fairly and is representative of real travel scenarios. Currently the levels associated with the on time attribute are as follows:

- 70% probability that the chosen travel mode will be on schedule
- 80% probability that the chosen travel mode will be on schedule
- 90% probability that the chosen travel mode will be on schedule

One potential conflict in the current survey design is the correspondence between the on time attribute with the intercity travel time. However, the proposed stated preference survey attempts to assess multiple level-of-service attributes and it can be assumed that the listed travel time is independent of the on time probability. In addition, the current on time attribute does not account for early arrivals, which is something that does occur in reality. For the proposed stated preference design, it was decided that the probability of delay would be considered. Depending on the results from the proposed study, future survey designs may assess the utility of including a set of levels that included both arriving on time as well as arriving early.

This level-of-service attribute is not expected to have a statistical influence on intercity mode choice. The prediction is that a respondent would not induce modal shifts if a preferred modal alternative has a slightly lower probability of being on time as long as the corresponding travel time is within reason. The on-time attribute is considered alternative specific attribute as there are only three defined and distinct levels.

### 3.2.6.6 Departures per Day

This attribute is named intuitively as it is the total number of departures that an intercity travel alternative would have within a given 24 hour period. The departure per day attribute is a modification of a headway time based attribute before the pilot test. Listing the total number of departures per day was preferred over a timed headway as some intercity travel mode alternatives do not follow an evenly spaced headway over the course of the day.

For the existing modes of non-automobile based travel, the number of departures in a day was determined by querying trips between Toronto and Montreal on three random days, of which one is a weekend. If the intercity mode alternative was provided by more than one service provider, then each provider was queried three times and the average of the total number of queries was
used as the baseline. In these multiple service provider situations, it was not necessary to coordinate the dates exactly as it is assumed that a departure per day is fairly consistent on similar days. It would be ideal to query the travel information for a specific day of travel; however, the limitation of the stated preference design procedure requires an aggregated baseline result.

As high speed rail is not currently in operation within the Quebec City – Windsor corridor, the number of departures per day was sourced from the 2010 EcoTrain report (EcoTrain, 2011), which stipulated a realistic projected frequency of travel to balance between ridership and operating costs. Compared to the existing intercity travel modes, the stipulated departure frequency was an acceptable assumption.

Once an averaged baseline departure per day value was calculated, the attributes presented to the respondent was a percentage of the baseline value rounded up to reduce any possible confusion from non-integer values. The associated departures per day levels are as follows:

- Baseline departures per day * 0.50
- Baseline departures per day * 1.00
- Baseline departures per day * 1.50

As daily departures ranged between a minimum of 8 and a maximum of 20 departures per day, the ±50% from baseline values is reasonable to assess how changes in the number of departure trips would affect intercity modal choice. The major assumption made in this departures per day attribute is that the variations in daily departures would not change the baseline cost of travel. The choice to keep variations in departures per day and travel cost separate is to separately assess how the two different attributes affect intercity mode choice. In addition, the level of variation chosen is not at an extreme value that would require significant changes in ridership costs to cover the difference in operational costs.

This attribute is not expected to be a primary contributor to a respondent’s intercity modal choice; however, there may be some statistically significant results when comparing different respondent demographics. For example, a business-based trip may place a high utility on more departures in a day while a personal-based trip may not be as sensitive to the number of
departures per day. The departures per day attribute should be considered a generic attribute even though the available departure values are fairly stratified.

3.2.6.7 Number of Transfers

This level-of-service attribute is defined as the number of times that the respondent would have to transfer from one transportation unit to another during the hypothetical trip from Toronto to Montreal. Transfers would be from one travel mode to the same travel mode. The local access and egress mode transfers do not count towards the value stipulated in this attribute. The purpose of including this attribute in the stated preference survey is to assess whether or not transfers influence a respondent’s intercity mode choice when explicitly stated as part of the intercity trip. Typical trips by non-automobile intercity modes are direct from origin to destination; however, this stated preference design adds in the scenario where transfers may be required when traveling within the Quebec City – Windsor corridor. The levels available for this attribute vary depending on the intercity travel alternative. For existing land-based travel alternatives, it was decided that there could be a maximum of two (2) transfers to assess whether the additional burden of transferring multiple times may affect the respondent’s intercity mode choice. For the air and hypothetical high speed rail travel modes, a maximum of one (1) transfer was deemed the upper limit that the modes should have given the travel distance between Toronto and Montreal. The following lists out the levels used in the number of transfer attribute as well as the applicable intercity travel alternatives.

- No Transfers – Bus, Rail, Air, High Speed Rail
- One (1) Transfer – Bus, Rail, Air, High Speed Rail
- Two (2) Transfers – Bus, Rail

The automobile mode does not include a number of transfers as automobile trips are expected to not consider transfers unless of adverse conditions or incidents. Like some of the other level-of-service attributes, they are not inherently tied to one another. For example, a trip alternative with two transfers is not necessarily tied with a cheaper travel cost or longer travel time. The absence of correlated attributes is to assess the significance of each attribute independently.

This attribute may contribute to a respondent’s intercity mode choice; however, the influence of this attribute may be dependent on the frequency that the respondent has traveled on the route as
well as any previous experience that the respondent has dealt with transfers when traveling local or non-local distances. The number of transfers attribute is considered alternative specific.

3.2.6.8 Seat Choice

The seat choice level-of-service attribute is defined as the availability of the respondent choosing his or her seat on an intercity trip. Unlike local travel mode options, there are currently no standing options available on Canadian intercity travel modes. While there is the possibility for a traveler to stand during travel, there is no added capacity of a travel unit due to standing ridership. The lack of a standing option on these travel mode alternatives is due to the long travel times in addition to safety concerns. Like the preceding level-of-service attributes, the purpose of including the seat choice is to assess whether or not the ability to choose seats influences a trip maker’s intercity modal choice.

In the early development stages of this proposed intercity survey, the seat choice attribute levels was related to the overall capacity of the travel mode. Capacity was defined as a function of available seats within a multiple of six seats. Under this type of level definition, travelers occupied from three (3) to six (6) of a multiple of six (6) seats. This type of level-of-service attribute was modified because it was assumed that the new methods of choosing seats would have a greater impact on intercity mode choice than the relative capacity of an intercity travel mode.

The increased use of online ticketing systems as well as electronic ticketing kiosks within travel stations has also increased the options for travelers to customize their own travel. For example, purchasing airplane tickets online also allows for a traveler to choose his or her seat at the time of ticket purchase or during check-in. It may be beneficial to assess how the availability (or lack of availability) may influence a respondent’s modal choices. The following lists out the levels used in the seating choice attribute.

- Pre-booked seating – seats are chosen at the time of booking
- Assigned seating – seats are assigned to traveler based on preference (i.e.: window or aisle)
- First come first serve – seats are filled at time of boarding
In the proposed stated preference survey, it is assumed that each intercity travel alternative may only offer the designated seat choice options despite those options not currently being offered. In addition, there may be other types of seat choice options available; however, to reduce complication to the stated preference design, the seat choice attribute has been kept to three (3) levels.

The seat choice attribute is not expected to have a significant impact on intercity modal choice. While it may be something that a trip maker considers, the respondent still has a seat during the trip. As the types of seat choices are stratified, this attribute is considered alternative specific.

3.2.6.9 Travel Information

Travel information is defined as the type of information that is provided for a potential trip maker about scheduling of the departing or arriving mode of travel. Before the large adoption of internet-accessible mobile devices and home internet access, departure and arrival information had to be accessed at the station. Currently, scheduling for many transportation modes are available on mobile devices and are dynamically changed if there are any deviations from the planned schedule. The following lists out the levels used in the travel information attribute.

- Mobile schedule – updated travel departure and arrival information is available on mobile devices
- Real-time schedule – updated departure and arrival information is available at stations
- Pre-posed schedule – no dynamically updated travel information, just a static daily schedule

The travel information attribute is not expected to have a significant impact on intercity mode choice. The difference between having dynamic schedule information and pre-posed scheduling is not expected to shift overall modal choices when compared to other level-of-service attributes. This attribute is considered alternative specific as the levels are non-continuous.

3.2.6.10 Intercity Travel Time

Intercity travel time is defined as the time required traveling the Greater Toronto Area to Montreal. The purpose of the intercity travel time attribute is to assess if the travel time for each travel mode alternative influences intercity modal choice.
For currently existing travel modes (excluding the automobile mode), baseline travel times were determined by using each travel mode’s ticketing website to query the estimated travel time for a number of different days. While travel times on weekdays should not differ to weekend travel times, both were selected and recorded. From the four or five recorded travel times for each modal alternative, the averaged travel time was used as a baseline value. There were three levels associated with intercity travel time and each of these levels was a multiple of the baseline intercity travel time determined from the averaged travel time. The associated access time levels are as follows:

- Baseline travel time (hours and minutes) * 0.70
- Baseline travel time (hours and minutes) * 1.00
- Baseline travel time (hours and minutes) * 1.30

For automobile trips, Google Maps presented an estimation of travel time based on perceived traffic volumes as well as an idealized travel time. The idealized automobile travel time was used as the baseline value. While automobile travel times are dependent on traffic and weather conditions, it was assumed that the coefficients used in the access time levels would account for any changes in automobile travel time. Another assumption was to assume that automobile trips have the same travel time from anywhere within the GTA. Omitting travel time within the city may not reflect the time needed to cross city border; however, it was assumed that the driving time to reach Montreal is a close approximation of all automobile-based trips originating from the GTA.

Currently available non-automobile intercity travel times are assumed to be from gate-to-gate as was stipulated from each modal alternative’s website. An assumption was made to not account for the time in the two scenarios below:

- Arriving at departure station and departing for Montreal
- Arriving at Montreal and existing the arrival station

This extra wait time was not included as it is variable depending on the time of the year, time of the day, whether or not the respondent has check in luggage, time spent at security checkpoints, etc. The large variability in the wait time is the main reason why it was not included within the stated preference survey.
For high speed rail, the baseline intercity travel time was sourced from the EcoTrain report (EcoTrain, 2011). As the EcoTrain report is the most recent comprehensive feasibility report regarding a high speed rail route along the Quebec City – Windsor corridor, the resulting travel time is considered accurate and representative of reality when compared to the other intercity travel mode alternatives.

The contribution of intercity travel time is expected to be a significant determinant of intercity modal choice. Intercity travel time would be considered as a generic attribute as access time is based on each modal alternative.

3.2.6.11 Intercity Travel Cost

The intercity travel cost attribute is defined as the cost of travelling from the Greater Toronto Area to Montreal but not including any costs associated with access or egress from the selected travel mode. This attribute was included into the stated preference survey because cost is hypothesized as one of the major factors that influence local modal choice. It may be beneficial to assess the importance that respondents place travel cost over other level-of-service attributes, such as travel time.

For currently existing travel modes (excluding the automobile mode), baseline travel costs were determined by using each travel mode’s ticketing website to query the estimated travel cost for a number of different days. Unlike travel time, travel costs do vary depending on the time of year as well as the day of the week. Travelling near holidays or weekends typically increase the cost of travel. In the proposed stated preference survey, trip queries were made for both weekday and weekend departures; however, major holidays were avoided. From the four or five recorded travel times for each modal alternative, the averaged travel cost was used as a baseline value. There were four levels associated with intercity travel cost and each of these levels was a multiple of the baseline intercity travel cost determined from the average weekday and weekend travel cost. The associated access time levels are as follows:

- Baseline travel cost (dollars) * 0.75
- Baseline travel cost (dollars) * 1.00
- Baseline travel cost (dollars) * 1.25
- Baseline travel cost (dollars) * 1.50
As detailed in 3.2.4.1, the cost of travelling by automobile is a function of the travel distance with an average cost of travel as stipulated by the Canadian Automobile Association. Similar to the assumptions made with the travel time, the distance of intercity automobile travel is the same for all respondents, which is a simplification in the stated preference design. Despite this simplification, the real-life travel cost is not expected to vary greatly from the baseline cost and any deviation may be covered by the four (4) levels stated above.

For high speed rail, the baseline intercity travel cost was sourced from the EcoTrain report (EcoTrain, 2011). As the EcoTrain report is the most recent comprehensive feasibility report regarding a high speed rail route along the Quebec City – Windsor corridor, the resulting travel cost is considered accurate and representative of reality when compared to the other intercity travel mode alternatives.

The contribution of intercity travel cost is expected to be a significant determinant of intercity modal choice. Intercity travel cost would be considered as a generic attribute as access time is based on each modal alternative.

3.2.7 Preliminary Utility Function

Given the above described attributes and levels, the following utility functions will be used as a basis for the SP survey design in the Ngene software. A general equation is used to describe all alternatives; however, it should be noted that the automobile mode does not have common-carrier attributes such as access/egress mode, transfers, and etc. This utility function does not represent the final utility equations for intercity mode choice; however, it is inclusive of all potential explanatory variables that will be in the final model estimations.

\[ Utility_{Alt} = \beta_1 \times Access\ Mode + \beta_2 \times Access\ Time + \beta_3 \times Egress\ Mode + \beta_4 \times Egress\ Time + \beta_5 \times \%\ On\ Time + \beta_6 \times Frequency + \beta_7 \times Transfer + \beta_8 \]
\[ + \beta_9 \times Seat\ Choice + \beta_{10} \times Trip\ Info + \beta_{11} \times Travel\ Time + \beta_{12} \times Travel\ Cost + \beta_{12} \times Premium\ Travel\ Time \]
3.2.8 Ngene Software

Ngene is a computer program used for generating experimental designs that are used in stated choice experiments (ChoiceMetrics, 2012). Given the input of the general utility function in the specified syntax, Ngene can be programmed to generate a basic stated preference survey skeleton that is useful for estimating choice models. The main benefit of using Ngene is the specific generation of orthogonal designs that are optimized to maximize effectiveness of survey result while maintaining orthogonality either across or within alternatives.

3.2.8.1 Full Factorial Design

A full factorial design considers all possible choice situation combinations for each attribute and attribute level. If a respondent was presented with a full factorial design, the collected responses would incorporate his or her choices for all potential permutations of the survey design. Collecting responses for the full spectrum of possibilities maximizes data analysis as no interpretation of missing scenarios is necessary (ChoiceMetrics, 2012). With a small number of attributes and attribute levels, this may be feasible without incurring survey fatigue. To calculate the number of possible scenarios in a full factorial design with $J$ alternatives, each with $K_j$ attributes, where attribute $k \in K_j$ has $l_{jk}$ levels, the following equation is used:

$$S^{ff} = \prod_{j=1}^{J} \prod_{k=1}^{K_j} l_{jk}$$

The current stated preference survey design includes; five (5) intercity travel mode alternatives and twelve (12) level-of-service attributes that affect at least two alternatives. Given the above list of alternatives, attributes and levels, there are over $1.2E25$ possible choice scenarios. While it would be mathematically preferable for each interviewer to choose an alternative for each scenario, it is deemed impractical to ask more than six (6) different scenarios to avoid survey fatigue.

The total list of possible full factorial scenarios, even when omitting contradictory attribute levels, will still be too large for respondents to attempt. Alternative methods of reducing the scenario size should be considered to collect relevant, detailed, and useful information without inducing survey fatigue on the respondent.
3.2.8.2 Orthogonal Design

A design is said to be orthogonal if it satisfies attribute level balance and all parameters are independently estimable. For orthogonal design, attribute levels for each attribute column in the design needs to be uncorrelated, resulting in the property that the sum of the inner product of any two columns is zero. Determination of orthogonal designs is a mathematically complex task that is further complicated when mixing different numbers of levels. As a result, computer programs often try to find near-orthogonal designs that can be used (ChoiceMetrics, 2012).

The orthogonal design allows for independent estimation of attribute influence on choice. Orthogonal designs are relatively easy to construct using computer software and maximize the ability of the model estimation efforts to detect statistically significant relationships.

In the orthogonal design of five attributes having three levels each, 18 choice situations are required for a true orthogonal design. These 18 scenarios may be too large to give to a single respondent without risking survey fatigue. To mitigate against fatigue, blocking can be used to split the orthogonal design into smaller subsets where the combination of all blocks is orthogonal. In Ngene, blocks are designed to present respondents with a level balance of attribute levels to avoid presenting respondents with only low or high levels (ChoiceMetrics, 2012).

At this stage of the survey design and implementation, an orthogonal stated preference design was used for the previously mentioned alternatives, attributes, and attribute levels. It is understood that only main effects are considered with an orthogonal design verses the additional interaction effects (between each attribute) of efficient survey design. The input code within Ngene’s software for simultaneous orthogonal design is detailed in Appendix D.

Using the Ngene code, six (6) different scenarios are produced by the Ngene software. These scenarios are the ones that are mathematically optimized to give the most statistically significant results. Given the large number of attributes included in the stated preference design, one of the potential limitations is that the smaller subset of six (6) scenarios may not be enough to assess how the attributes are related with mode choice. One solution would be to design a larger number of scenarios within Ngene and to create blocks of six (6) scenarios within the survey; however, the low number of expected survey responses may result in an underrepresentation of
each scenario. In addition, another barrier to having more than six (6) stated preference scenarios relates with the web-based survey instrument and is further detailed in section 5.2.2.

3.2.8.3 Efficient Design

Efficient design was not incorporated into the survey design and build; however, the premise of efficient design should be considered for possible future expansion this or similar projects. The results of model estimation in this project could be used as a platform for efficient design (ChoiceMetrics, 2012).

The goal of efficient design is to obtain data that generates parameter estimates with the smallest possible standard error. There is a catch-22 with the application of efficient design in stated preference survey design as parameter values are required to estimate parameter values. To mitigate this contradiction, prior parameter estimates from literature or similar studies could be used to determine the asymptotic variance-covariance matrix. One potential flaw of using prior values is the assumption that these prior parameter estimates are correct and match the attributes of the proposed survey.

3.3 Local Revealed Preference

In addition to the intercity travel revealed preference questions, a set of local travel revealed preference questions was also included within the survey design. The main reason that this set of questions was included in the proposed survey was to have comparable travel data in the event that the respondent has not made an intercity trip within the Quebec City/Windsor corridor within the past year. During survey development, there was a concern that the addition of a local revealed preference survey would induce survey fatigue. In the proposed survey design, this set of questions was placed immediately after the completion of the stated preference section. It was feared that prolonging the survey beyond the stated preference questions would cause a respondent to not fully complete the survey. However, compared to the alternative of placing the stated preference survey near the end, it was deemed more important to collect the SP information and risk having incomplete local RP data.

The main purpose of this question set is to collect the respondent’s daily travel and activity information. The first question asked in this set is the respondent’s daily travel purpose and deviates into three potential question sets depending on the chosen response. If the respondent
states a daily work commute, additional questions regarding work location will be asked. Similarly, if a non-work and non-school mode is chosen, the type and location of typical daily travel will be asked. Location of school-based trips was not asked in case of any privacy issues.

Following the purpose-specific questions, general questions regarding mode choices, typical travel time, and typical travel costs are asked. These questions are typical of other local daily travel based surveys with the option to select multiple travel modes as well as stratified travel cost and time options. Stratified travel cost and time was chosen under the assumption that trip makers typically assess his or her travel cost and time in an averaged or rounded value than a specific unit metric.

The collection of this data may potentially be used to correlate a respondent’s daily travel patterns with his or her intercity modal choices. Potential correlation could be determined by entering relevant revealed preference answers into the utility function of intercity mode choice. However, the main purpose of the local revealed preference data is to cross reference the demographics of the respondents from this survey with demographics from other travel data sets; such as the ones from Transportation Tomorrow Survey (TTS) from University of Toronto’s Data Management Group. As the TTS collects a large number of respondents from a known sample listing, cross referencing the collected responses from the proposed intercity survey would validate the demographic of collected data.

### 3.4 Demographics

This final section of the proposed survey design inquires about the respondent’s individual and household socioeconomic information. As the proposed survey does not require specific information about a respondent’s name or home address, the collected data remains anonymous. Available answers for the socioeconomic questions are either specific (household size) or is between a range of values (age or income). The range of questions included are to inquire about the respondent’s demographic as well as household demographics that have been historically shown to influence transportation-related choices. While the included questions is not an exhaustive list of socioeconomic questions, a balance was required between collecting relevant data and mitigating survey fatigue, especially after the previous three sets of questions.
Similar to existing travel surveys, the final set of questions asks some basic socioeconomic questions for two main reasons. The first reason is to establish a proper socioeconomic profile when compared to previously collected data sets; such as the Canadian census or the Greater Toronto and Hamilton Area’s Transportation Tomorrow Survey. By comparing the collected socioeconomic data with existing data sets, it is possible to be more confident that the collected data is representative of the target population frame.

The second reason is to correlate the collected revealed preference and stated preference data with an individual’s socioeconomic level. Combining the two aspects may potentially allow us to analyze how an individual’s household income or household car ownership may affect the choices that he or she may make when making an intercity trip.

One potential issue is the problem where a respondent would not complete the socioeconomic questions and thus providing an incomplete survey data set. An incomplete data set is defined when the respondent does not complete all the mandatory questions within the survey instrument. While an incomplete could still be used for modelling purposes, it is ideal to have a completed data set to ensure that the information provided is valid and relevant. The set of questions deemed most likely for the respondent to prematurely leave the survey is the socioeconomic questions because it is at the end of the survey and may ask about household information that the respondent may not be willing to openly disclose, despite anonymity of the survey instrument.

To mitigate against this problem, the socioeconomic questions have an “undisclosed” option in case a respondent does not feel comfortable with providing an answer. This “undisclosed” option allows for the respondent to not state a specific value for questions that they do not feel comfortable; however, allows for the survey to progress to the next question. The one disadvantage would be if there were too many “undisclosed” answers reported by respondents as this option does not provide data on a respondent that is useful for statistical analysis.
The chapter explores how some of the more intricate details were incorporated into the survey design. These details include the disaggregation of geographical locations, to provide specific local travel times for respondents, as well as the procedure to output a large number of stated preference tables while minimizing human error. In addition, these details were integral to the design of the proposed survey and had to be completed using computer software or time intensive labor.

Several main computer programs and services will be further explored to explore the process and procedures used in creating these elements of the proposed survey. One of the objectives is to use readily available programs and services; as a result, mainly Excel and Google Maps was used. The goal was to show that software used industry-wide is capable at producing complex functions. While there are inherent limitations, it was possible to overcome most of these limitations by intuitive problem solving.

The proceeding discussion will explore these elements in the context of the proposed survey; however, the logic and procedures used may be applicable to various other research topics. Three main survey elements will be explored; referencing FSA locations, using Google Maps to find travel times, using Excel to create the stated preference tables, and analyzing collected data.

### 4.1 Forward Sortation Area

Access and egress times were two of the attributes that were hypothesized to influence intercity modal choice. This section details how origin locations in the Greater Toronto Area were disaggregated down to the Forward Sortation Areas (FSA) level. As discussed in section 3.2.6.1, local egress destinations were not disaggregated to this detail based on the assumption that a respondent typically associates with a general destination area rather than a destination postal code. Given the size of the GTA, the main challenge was to accurately account for all possible FSA’s and to provide accurate information to the respondent.
Rather than providing a single central metropolitan area based trip travel time, the project scope required a high level of local disaggregation for trip origins. In a balance between computational complexity and level of detail provided to the respondent, it was decided that aggregating respondents by the FSA level would provide satisfactory detail regarding local access within reasonable computational requirements.

From a commonly used six-character postal code, the Forward Sortation Area is the first three characters, which is generally used as the basis of sorting mail. The first character identifies one of the 18 major geographic areas, provinces, or districts in Canada (M represents metropolitan Toronto and L represents central Ontario). The second character is a numerical digit that identifies if the area is urban (by numerals 1 to 9) or if the area is rural (by the numeral 0). Finally, the third character designates the forward sortation area segment in conjunction with the first two characters (Canada Post, 2013). Currently, there are 205 FSAs spread throughout the GTA. The shape and size of these FSAs are roughly based on population density and geographical constraints; however, it also takes into account businesses and other non-residential addresses.

4.1.1 Problem

In order for travel times from FSA origins to be found, it was necessary to either have a location on a map or the longitude and latitude attributes of each FSA. A Google Map based platform was preferred as Google Maps provided travel times based on a number of various modes. Additionally, the customization options on Google Maps would allow for the modification of an existing map based on requirement. For example, if a map of FSA boundaries was available, the FSA centroid could be tagged with a place marker and used to query travel times.

Initial internet queries on existing non-GIS based maps indicating FSA boundaries were unsuccessful. While PDF maps of FSA areas were available, it was time intensive to trace the boundaries of each individual FSA. As seen in Appendix E, one of the more detailed maps was created by Canada Post where each FSA is listed alongside the boundary roads; however, the irregular geometry of many FSA’s would make the hand-drawn process too time intensive and potentially inducing human errors.
One alternative was to utilize available GIS-based boundary files from Statistics Canada (Statistics Canada, 2013). Having the GIS information, it would be possible to import the data into software, such as ArcGIS, and create a centroid from the vertices of the FSA shape. Reasoning against using GIS data is the limited use of GIS-based software compared with more industry-wide software such as Microsoft Excel. In hindsight the use of the Statistics Canada FSA boundary file with ArcGIS may have simplified this process of creating FSA centroids.

4.1.2 Data Source

In the absence of a Google Maps based map of FSAs further online searches lead the company Geolytica Inc., operating out of Findlay Creek in Ottawa, Ontario under the website GeoCoder.ca. On the GeoCoder.ca website, a Canadian postal codes dataset is available under Open Database License (ODbL). This dataset is available for free in addition to a paid version that includes street names and address ranges associated with each postal code for commercial and marketing purposes (Geolytica Inc., 2012). This dataset has been crowd-sourced since 2004 and the information has been derived from open data sources with good accuracy at the postal code level.

While this Geolytica Inc. data set may not be as robust as the set provided by Canada Post, the postal code file from Canada Post is not open source and costs approximately $5,000 CAD for access (Geolytica Inc., 2012).

The most current dataset by Geolytica Inc., updated on March 1, 2013, contained 927,628 unique postal codes in Canada, verified and geocoded to street level accuracy. The postal code area ranges from a single house, city block, or whole town depending on population and development density. To obtain the geographical representation of a postal code, the geodesic point (a point with the smallest relative distance from all locations in the postal code area) is found.

A free Canadian postal code data set from GeoCoder.ca was downloaded on September 26, 2012 to obtain FSA centroid locations. This data set contained 914,084 individual postal code records with the corresponding longitude, latitude, province, and city. To confirm the validity of the data set, the longitude and latitude of known postal codes were queried on Google Maps to verify location. Table 4-1 below lists the postal codes used for verification purposes.
With the scope of the proposed project being limited to the Greater Toronto Area, it was beneficial to filter the total data set to only display postal codes within the project scope. The benefit of this GeoCoder.ca data set was the province and city attribute. First non-Ontario postal codes were filtered resulting in a total of 166,498 postal codes documented in Ontario. Next, postal codes not starting with the letters M (for Metropolitan Toronto) and L (for Central Ontario) were filtered, further reducing the total number of postal codes to 132,230. However, the number of postal codes within the project scope is still too large to individually obtain local access travel times; as a result, it is necessary to aggregate these six-character postal codes into the three-character forward sortation area.

### 4.1.3 Postal Code versus FSA

In the decision to aggregate household origins at the forward sortation area (FSA) level, it was necessary to conglomerate the existing six-character postal code coordinates into a representative single aggregated three-character FSA coordinate. The main reason to aggregate is to simplify required calculations while retaining relevant local travel times. The main technique used in aggregation is to calculate a geographic midpoint based on available postal code data using the industry-wide software Microsoft Excel.

There are a number of calculation methods that could be used to calculate the FSA midpoint; geographic midpoint, center of minimum distance, and average latitude/longitude. A geographic midpoint is the average coordinate for a set of points on a spherical surface. The center of minimum distance locates the coordinate that minimizes the combined travel distance from a set of points. While this method obtains a point of absolute minimum travel distance, it does not equalize the distance traveled from non-minimum points. An average latitude/longitude is the midpoint of multiple points on a flat project map. In areas where the spread of the points spans less than 400km, then this averaging method outputs close approximation as the geographic midpoint method (GeoMidpoint, 2012). A detailed guide on calculation methods as provided by

<table>
<thead>
<tr>
<th>Postal Code</th>
<th>Reference Location</th>
<th>Confirmed</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>T5T 6B1</td>
<td>Previous Home Address</td>
<td>Yes</td>
<td>30m</td>
</tr>
<tr>
<td>M5T 1R1</td>
<td>89 Chestnut Residence</td>
<td>Yes</td>
<td>10m</td>
</tr>
<tr>
<td>M5V 3Y3</td>
<td>Current Home Address</td>
<td>Yes</td>
<td>10m</td>
</tr>
</tbody>
</table>

*Table 4-1 – Verified Postal Codes From geocoder.ca Dataset*
GeoMidpoint is located in Appendix F. Given the relatively small spread of points for each FSA calculation, the method of averaging latitude/longitude was used.

After filtering down the raw data from the GeoCoder.ca website down to the 132,230 postal code located within Central Ontario and Toronto, the procedure to calculate averaged latitude and longitude coordinates is further described in Appendix G.

For example, for the FSA designation M5T, located in downtown Toronto, there were a total of 454 postal code records with the same FSA designation. By using the procedures in Excel, the geographic midpoint was determined to be 43.65351862 latitude and -79.39749533. As seen in Figure 4-1 below, a search of that specific coordinate in Google Maps is a point on the northeast corner of Dundas Street West and Spadina Ave. Referencing with the 2001 postal code map provided by Canada Post, in Figure 4-2 below, the M5T FSA is bounded by College Street to the north, University Ave to the east, Queen Street West to the south, and Bathurst Street to the west. The coordinates computed in Excel is located roughly in the central of the FSA boundaries, which is an indication that the procedures used can accurately determine FSA centroids.
4.1.4 Problems and Limitations

Compared to finding a geographical centroid using the vertices of the FSA boundary, the proposed method of finding the geographical midpoint would account for the varying population density within each FSA. As many geographical coordinates are used in the calculation of the geographic midpoint, there is a reduced risk that an erroneous data point would greatly affect the midpoint calculation. One potential downside would be the influence of non-residential postal codes in the calculation the geographic midpoint; however, the difference between an overall density-based geographic FSA midpoint and a residential origin based geographical FSA midpoint would be minimal despite the increased data and time required to filter out non-residential postal codes. In addition, the alternative method of calculating geographic midpoints by FSA boundary vertices would also induce the error of accounting for non-residential origins.

Another possible limitation to this approach is the data set itself. As the records on this data source are crowd-sourced and freely available, there is the potential of some errors. These errors may be typing mistakes in the database or missing records. As Canada Post’s own dataset is not publically available, it is not possible to compare between the two data sources. However, the large number of records available in the data set would possible negate the errors induced the few errors. Having 1000 erroneous records within the Central Ontario section would still be under 1% of all records. Additionally, the aggregation of a large number of individual postal code records to the FSA level would reduce the effect of errors on the validity of calculations.
4.2 Google Maps

After creation of the FSA centroid coordinates, the next step is to utilize the obtained coordinates in a suitable software/service for further use and analysis. To reiterate, the main purpose is to provide disaggregated local access travel information to infer any relation between local accessibility and intercity mode choice. With the list and coordinates of FSAs in the Greater Toronto Area, the next step would be obtaining local travel times with various local travel mode options. To satisfy this requirement, a computer program or online service should be used to query local travel distances and travel times.

This computer program or online service should be able to provide trip information beyond door-to-door travel distance. While many electronic map services may provide door-to-door distances, there are only a handful of online services, such as Google Maps and Bing, which provides scheduled transit or non-motorized trip information. In the case of Google Maps and Bing, transit operators provide relatively updated schedules and integrate the respective transit system map into the map interface. In addition, another non-essential requirement is the ease of integration between the obtained FSA coordinates into the mapping interface. A visual feedback of the obtained FSA coordinates would be useful in verification of the coordinate creation process as well as visual feedback of the potential routes that would be taken.

Based on the above mentioned requirements, Google Maps was used because of the proven accuracy of the mapping interface as well as the integrated application programming interface (API) that is used to incorporate external data, such as transit routes and schedules, into the main user interface. Additionally, it is also possible to create custom maps that indicate certain locations or areas in Google Maps. The following sections will discuss about the benefits and limitations of using Google Maps as well as the procedures used to query local travel times in the context of the proposed project.

4.2.1 Adapting FSA Coordinates

One of the main reasons why Google Maps was used as the online mapping service over competing services was the ability for users to easily create custom maps. In these customizable maps, it is possible for the map creator to place various points of interest using various marker styles and attribute those points of interest with defined names and descriptions. The objective is
to define each FSA coordinate as a marker and use Google Maps’ direction function to obtain travel times from each FSA centroid to the necessary destinations.

It is possible to copy and paste the coordinates of each FSA centroid into Google Maps’ search function and then save that specific location with a location marker; however, given the large number of FSAs, a more automated method was preferred to streamline the points creation process. When creating a custom map, there is an option to import map data from KML, KMZ, or GeoRSS files into the map. For the scope of this project, only KML files will be discussed. KML is an open standard officially named as OpenGIS® KML Encoding Standard (Google, 2013) and contains information with attributes related to spatial location and description of map-based objects. As the calculated FSA coordinates are not in the KML format, it is necessary to find some conversion tool in order to properly import and display the FSA coordinates in Google Maps.

To create KML files, Google has developed a spreadsheet-based program that creates a KML document for import under the interface of a spreadsheet document. This program is called Spreadsheet Mapper and version 3.0 was used in the project. At the time of writing, version 3.1 is available. It should be noted that Spreadsheet Mapper is only available for use on Google’s own office suite under the Google Drive name as it utilizes Google Apps Scripts in order to convert data into KML files (Google, 2012). Google has provided an easy to follow guide on how to use the Spreadsheet Mapper 3 program (Google, 2012) and the available options within the program are more complex than the requirements in the project scope.

The basis of creating an importable KML file into Google Maps is to copy and paste relevant columns of information into the Spreadsheet Mapper 3 program\(^2\). The first sheet, labeled “start here”, is used to place basic information including the name of the creator and name of the KML document for referencing. The second sheet, labeled “PlacemarkData”, is where the FSA coordinates are copied into. The columns used in this PlacemarkData sheet are as follows:

\(^2\) To facilitate the movement of data with minimal errors, the columns in the Excel spreadsheet of the FSA coordinates should be rearranged to the same order as the Spreadsheet Mapper 3 software. If there are no errors in the PlacemarkData sheet, as indicated by the Error Check Results column, then the KML file can be downloaded by a link available on the “start here” sheet in Spreadsheet Mapper 3. With the KML file downloaded, this file can then be imported into Google Maps with each FSA centroid point designated by an individual marker.
- Folder Name column is used to designate the corresponding city for each FSA
- Placemark Name column is the FSA itself
- Coordinates are intuitively the corresponding latitude and longitude for each FSA
- Template is set at Template1 for all points and will output the same type/color of marker in Google Maps. If different marker styles are desired, then multiple templates can be used.

![Screenshot of Spreadsheet Mapper 3 layout](image)

Another function of Google Maps is the ability to draw polygonal shapes and lines directly onto the map. In this project, various polygons were drawn to indicate the various regions that comprise the Greater Toronto Area. Lines were added on top of the shapes to create the political boundaries between the various cities. Displayed below in Figure 4-4, overlying the FSA points on top of the boundary shapes makes it easier to identify each city’s corresponding FSAs and these can be verified by cross-referencing with existing maps. Drawing of the boundaries also makes it easier to eliminate FSAs that are beyond the GTA and were erroneously included in previous steps.
4.2.2 Query Travel Data

With the FSA centroid points imported into Google Maps and displayed as individual markers, it is possible to query trips from each FSA centroid point into various destinations. The basic procedure to look up a destination is to access the custom map with the FSAs and go down the sidebar list. It was found that going down the sidebar list is a more systematic approach compared to drawing from the markers on the map. When a marker is selected, a speech bubble appears with the option to find directions to that location. Based on the purpose of documenting travel time from each FSA to relevant intercity departure stations, the following destinations were used as destination locations.

- “Toronto Coach Terminal” for access to the Greyhound and Megabus departure station
- “Toronto Union Station” for access to Via Rail and high speed rail departure station
- “Toronto Pearson International Airport” for access to Air Canada and Westjet departures

Based on the 205 FSAs and three (3) possible destinations, 615 individual queries were made and recorded on a new Excel spreadsheet. For each OD pair, the automobile trip distance, automobile travel time, transit travel time, and walking travel time were recorded. Biking trips were not recorded as Google Maps’ bicycling directions were still in beta mode. The following paragraphs explain the assumptions made when documenting travel time for each local mode.

For automobile trips, the shortest route to the destination was recorded. Typically the longer alternative routes suggested by Google Maps had similar travel times that were often within five (5) minutes of the suggested route with the shortest travel time. In addition, if Google Maps
indicated that the suggested route has tolls, it was assumed that the respondent would still use those roads without the additional price increase of the toll. As the price of local travel was a range of prices dependent on the travel mode, this simplification is not expected to greatly affect the survey results.

For transit based trips, Google Maps has travel options for a number of transit operators including; TTC, MiWay, GO, Brampton Transit, and York Regional Transit/Viva. When a suggested transit travel route, which is typically the shortest travel time, involved a transfer between transit operators, a route with a longer travel time but with no transfer between transit operators was chosen instead. This assumption was made to replicate realistic travel choices in the situation where an individual would be using transit to access a departure station. In addition, all transit-based trip queries were made for a 13:00 departure time on a weekday. The standardization of departure time was made to minimize variation of transit service during the course of the day.

For walking based trips, similar assumptions were made as automobile based trips. The shortest route was recorded. If a walking trip had travel time beyond an hour, the travel time was still recorded and will be later filtered out in the Excel stage of the survey build.

In hindsight, it would have been beneficial to also query trips from each FSA centroid to closest entrance into Montreal. Given the large geographical spread of the GTA, it was only assumed that automobile trips to Montreal are from downtown to downtown. This assumption negates the additional travel time induced by travelling through the GTA to access the right city exit.

### 4.2.3 Benefits

The primary benefit of utilizing Google Maps as the data source is the intuitive nature of obtaining relevant travel times and distances. The constant development from Google at improving the direction routing algorithm as well as the integration of multiple transit routes and schedules from a number of transit operators in the GTA were helpful to collect data. With the further integration of the custom created FSA centroid points, it was a simple process of a few computer clicks to search for a suggested travel route. The intuitive nature of Google Maps reduces the necessity to go through background databases. For example, the Toronto Transit Commission’s routes and schedules are available as open data (City of Toronto, 2013); however,
additional effort would be required in order to distill the data. Additionally, the integration of multiple transit operators into the background database in Google Maps allows for the transfer between multiple transit alternatives where necessary. The ability to assess travel through multiple transit operators with the correct routes and travel times would be a large extraneous task if not for the integration within Google Maps.

4.2.4 Limitations

While there are tools available online to batch queries into an automated process, Google Maps has not included transit data within the accessible Google Maps API. As a result, it is only possible to query automobile trips but not transit based trips using automated computer software. In addition, Google has limited the number of automated queries to 2,500 a day (Google, 2013), which is more than the current project scope, but may be a limitation if the scope expands. Future iterations of this project could compare travel times and distances through batched searches with the currently used manual input method.

Another limitation experienced when using Google Maps to obtain local travel times was the necessity to trust the validity of the data. Especially with transit travel time, while the listed travel time is an approximate based on transit scheduling, the reality of traveling by transit often incurs a higher travel time because of traffic conditions, weather conditions, delays in scheduling. Without another easily attainable source of information, the information provided by Google Maps had to be trusted.

4.3 Excel

Microsoft Excel, known further as Excel, is a database software introduced primarily for accounting purposes but continued development and refinement of the program has created additional functions useful for applications where a large amount of data and/or calculations are required. In the field of science and engineering, there are other software alternatives such as MatLab, Maple, Mathematica, etc. that were designed for more sophisticated mathematical applications (Liengme, 2000); however, the scope of the proposed project was anticipated to have a large number of simple calculations spread over a number of separate documents. Based on the scope and the availability of the numerous software alternatives, Excel was the alternative that satisfies the possible data management requirements. The following sections discusses the
4.3.1 Rational

The rationale behind using Excel as the primary data management software was the widespread use of Excel within both the engineering and marketing industries. Referring to one of the main project objectives of creating a complex data collection tool using industry-wide software and web services, Excel is currently used in many industries with many individuals understanding the basic structure of data entry, data modification, and data management in Excel. The benefit of the widespread use of Excel across different industries is the recognition of common elements within a spreadsheet. For example, data collected by a marketing research agency could be easily interpreted and analyzed by an engineer under the Excel interface.

There are numerous functions in Excel and many individuals and businesses may use a fraction of the available data management and analysis tools built into Excel, which is both an advantage and disadvantage to the use of Excel. The advantage is the idea that there is a large potential for one program, Excel, to manage a set of data. In addition to the simple data management, various functions may be used to output relevant information required for analysis and data modelling. The main disadvantage to this is the number of individuals who can utilize all the lesser-used functions in Excel compared to the number of individuals who only use Excel for basic tasks. Under the premise of industry-wide use, the software and basic elements may be used industry-wide; however, advanced features and logic may add confusion for basic users. In the goal to simplify various aspects in the survey design and build, a more complex use of Excel may be a barrier to widespread adaptation. To mitigate this barrier, there are some possible procedures and tools that can be used including: documentation of Excel use, labeling of elements within the Excel spreadsheets, and creating limitations within Excel to prevent misuse.

With the numerous file types available for use, it is important for software to support the import and export of a number of file types with minimal work required to adapt the files for use. Excel is one such program that can readily adapt some of the most commonly use file types including but not inclusive of; Microsoft Access, web documents, text (.txt and .csv) files, SQL server table, and XML files. Text files, especially comma-separated values (.csv), are simple file formats that are supported in different industries such as consumer, business, and scientific
applications (Wikipedia, 2013). While the ability to import and export data sets are not exclusive to Excel, the functionality of import data and the available options in Excel is satisfactory within the scope of the project.

In the survey design, large computations are required to output the necessary files. In creating the stated preference tables, the data from a number of spreadsheets are combined together to output a specific origin-destination pair. The use of Excel is beneficial to completing the required computations without significant error that may cause program crashes. It should be noted that performance of Excel is also a factor of the computer hardware running the Excel program, where older machines may experience trouble in handling the Excel files created in this project. The full extent of the calculations required under the project scope are documented in the sections below where the computations required are discussed and presented with relevant flow charts if applicable.

The learning curve in Excel may depend on the complexity of the data requirements in the spreadsheet and may also depend on the previous experience of the user. One of the benefits of Excel is the availability of resources, especially on the internet, regarding the use and functionality of Excel. Because of the large user base from many industries, a large pool of information is available. This information may reduce the learning curve in Excel for individuals who seek out these additional resources. During the course of this proposed project, these external information sources have been a benefit of introducing new functionality and work-around in data management. At times, multiple sources on the same topic were available and the combination of the multiple sources was instrumental at understanding the form and function required. Lastly, a large majority of this information is publically available on the internet and is readily available.

4.3.2 Stated Preference Setup

This section details the use of Excel when creating the tables for the stated preference section of questions. Some details in this section reference back to the various level-of-service attributes in section 3.2.6. Each sub section relates to the individual sheet in the Excel spreadsheet file. The main idea of this setup is disaggregating each element of the stated preference design into separate sheets and then combining the various elements into a defined output form.
The following sub-sections outlines the purpose and function of each sheet as well as the Excel formulas used. When the Excel formulas are explained in detail, a designation\(^3\) will be used to indicate the referenced cell(s) followed by the formula and a description of the function. A facsimile of the Excel sheets are located in Appendix H and specific parts of the spreadsheet may be presented in further detail and context where applicable. The general flow of the Excel spreadsheets when setting up the stated preference tables is detailed in Figure 4-5 below.

![Flow of Data in Microsoft Excel](image)

When discussing about the logic and formulas used in Excel to execute a function, it is possible that there are other methods that may achieve the same result. In the discussions below, only the methodology used will be discussed.

### 4.3.2.1 Variable

The *Variable* sheet lists the defined level-of-service attributes used in the stated preference (SP) survey, the corresponding levels for each level-of-service attribute, and the possible levels and attributes for each alternative. One purpose of this sheet is a quick visual reference to the level-of-service attributes and corresponding levels used in the SP design. The layout of this sheet is set similar to the final SP table only with all the possible levels visible. While corresponding

\(^3\) Cells will be referenced as [cell reference], for example column B, row 34 is designated as [B34]
levels for the different alternatives are the same, it is repeated in anticipation for a more complex SP survey design. Listing each attribute and level is useful to visually indicate differences, such as the lack of two (2) transfers for the air and high speed rail modal alternatives.

Another purpose for this sheet is to act as the contextual reference for the SP table\(^4\). In Ngene, the outputs provided by the software were either a numerical reference of the level (for attributes such as seating choice) or the specific value of the attribute (for attributes such as a coefficient of travel time). In the case of numerical reference of the level, this Variable sheet is used to prompt the display of the corresponding text string. This cross referencing is further explained in section 4.3.3.

In the case where the corresponding attribute levels are coefficient values, a custom number format is used to visually indicate the purpose of the cell. For example, for the total travel cost attribute, the corresponding levels (as explained in section 3.2.6) are percentages of baseline travel costs. This function is indicated as [D33] “Existing * coefficient”. In this case, a custom number format is used to attach the text “Existing *” as a virtual indicator and only the coefficient (in this case 1.00) is typed in and stored as a number. The purpose for this is to retain the number format of the cell when the SP table references the coefficient to calculate total travel cost\(^5\).

4.3.2.2 Access

The Access sheet is a database of local location and travel information used as inputs into the stated preference table. The rows of the Access sheet are based on each of the 205 FSAs in the Greater Toronto Area. The columns of the sheet refer to; the corresponding city, geographical coordinates, and relevant local travel times. The main purpose of this sheet is to store these local accessibility data for use in the stated preference survey.

\(^4\) A column [B:B] is used to match the order of levels in the Ngene SP design with the proper outputs in Excel, which is important to make sure Excel picks the proper level in the next steps.

\(^5\) The custom cell format is ["Existing * "0.00] where the text in between the quotes becomes a virtual text string and the 0.00 indicates an output of a numerical value with two significant digits.
A FSA identification number is created for each of the 205 FSAs, which are arranged in alphabetical order of the corresponding city name. This identification number is required to reference each FSA and the corresponding local accessibility information in the stated preference tables. The coordinates and corresponding city for each FSA are based on the work done in section 4.1. For the local access distance and time, the suggested driving distance, auto travel time, transit travel time, and walk travel time data from Google Maps to the three possible local destinations are listed. The procedures of obtaining these travel distances and times are detailed in section 4.2.

No additional calculations or data modification are required in the Access sheet as the majority of the work is completed prior to this stage.

4.3.2.3 Egress

Similar to the Access sheet, the Egress sheet is a database of local location and travel information used as inputs into the stated preference table. The rows of the Egress sheet are based on the 19 boroughs in Montreal. The columns of the sheet refer to: the corresponding city, geographical coordinates, and relevant local travel times. The main purpose of this sheet is to store these local accessibility data for use in the stated preference survey.

An identification number is created for each of the 19 boroughs, which are arranged in alphabetical order of the borough name. This identification number is required to reference each borough and the corresponding local accessibility information in the stated preference tables. The coordinates for each borough are based on the work done in section 4.1. For the local access distance and time, the suggested driving distance, auto travel time, transit travel time, and walk travel time data from Google Maps to the three possible local destinations are listed. The procedures of obtaining these travel distances and times are detailed in section 4.2.

No additional calculations or data modification are required in the Egress sheet as the majority of the work is completed prior to this stage.

4.3.2.4 Schedules

The Schedules sheet is used as an ongoing database to continuously collect information regarding trip attributes for non-automobile intercity travel modes. The main purpose is to obtain
a representative daily departure frequency, intercity travel time, and intercity travel cost for these modes. Currently, there are only 15 entries in this database; however, this can be expanded in the future to include trips to multiple destinations from multiple origins.

Under the current scope, the stated preference tables are generated before survey distribution. Because these tables are generated prior to distribution, it is not possible to include dynamic travel details such as the cost of intercity travel for a specific date. As a result, it was determined that an averaged value of travel attributes over a number of days would be the best representative of a travel for any day while incurring the least errors.

For the final survey used in this project scope to collect data, three travel times for each intercity travel mode was queried. For each travel date, the number of departures in that day, the average travel time of all departures, and the average travel cost of all departures were recorded. Currently, the averaged frequency, travel time, and travel cost have to be calculated by creating data filters; however, it is possible in future iterations to create a set of formulas to draw from relevant data points automatically.

4.3.2.5 Destination

This Destination sheet can be interpreted as the simplification of the Schedules sheet. The purpose of the Destination sheet is to display the baseline travel values used in the stated preference tables. The values on this sheet would be used as reference values on other sheets in the Excel file. Essentially, this sheet is a summation of the previous Schedules sheet with the addition of the automobile and high speed rail travel attributes. Currently, this spreadsheet is sparse on data as it only contains the baseline travel attributes with Montreal as a destination; however, the intended purpose of this sheet is to allow for possible future expansion with a multitude of different origin-destination pairs.

No additional calculations or data modification are required in the Destination sheet as the majority of the work is completed prior to this stage.

4.3.2.6 NgeneRAW

The NgeneRAW sheet is a direct import of the text output file from the Ngene software. In Ngene, a text file is created from the tables generated by the software and saved as a plain text
(.txt) document. In Excel the following steps were used to import the .txt file into the spreadsheet.

- Import data from text file and select the .txt file that the Ngene scenario tables were saved under.
- A pop-up window for the text importer wizard will appear. Under the “Original data type” choose delimited as each cell is separated by a non-fixed number of spaces.
- In step 2 of the text import wizard, choose space as the delimiter and check the box that treats consecutive delimiters as one. This step would look at the text file and put data on a new cell once there is more than one space between values.
- In step 3 of the text import wizard, the column data format can be left as generic and the “finish” radio button can be pressed to complete the data import.

Some initial observations are that the order of the attributes is not correctly aligned with the right alternatives. This misalignment occurs because the import process considered multiple spaces as a single delimiter. In addition, the attributes are arranged by the number of modal alternatives that each attribute is correspondent with. For example, premium travel cost is an attribute of rail, air, and high speed rail and is placed on the very bottom while travel time is an attribute for all modal alternatives and is placed on the top. Under these two observations, the attributes were rearranged to justify to the right side (rather than the left) to correlate with the initial stated preference design. The Variable sheet may be used for cross reference.

No additional calculations or data modification are required in the NgeneRAW sheet as the majority of the work is completed prior to this stage.

4.3.2.7 Ngene

The Ngene sheet is the NgeneRAW sheet but the order of the attributes has been arranged to the order that the attributes will be presented in the stated preference survey. To reduce on the potential for error, reference cells are used instead of copy and pasting data. As each of the six

---

6 Access Mode for SP Scenario 1 was in cell [C6] on the NgeneRAW sheet but is moved to cell [C5] on the Ngene sheet. If cell [C5] on the Ngene sheet is copied to Access Mode for SP Scenario 4, cell [J5] on the Ngene sheet, then Excel would grab the data from cell [J6]. Figure 4-6 below shows the change between the NgeneRAW and Ngene sheets.
scenarios is arranged in the same way, it is possible to copy the formulas on one SP scenario and paste for the remaining five. Based on the way Excel uses copy and paste, the copy function will copy the formula and remember the location of the copied cell and when pasted, a new reference cell would be used the same distance away from the original cell.

![Figure 4-6 – Comparison between NgeneRAW and Ngene](image)

No additional calculations or data modification are required in the Ngene sheet as the majority of the work is completed prior to this stage.

4.3.2.8 SP Single

The SP Single sheet is the primary spreadsheets that aggregate the previous information into the stated preference design. As the name of the sheet implies, the SP Single sheet refers to one of the 23,370 possible stated preference combinations that are required in the scope of this project. The main purpose of this sheet is to draw the relevant information from the previous sheets into one location. It was found that combing the data into the stated preference table was easier when there was an intermediate sheet that displays the information for a specific origin destination combination.

The following subsections discuss the various elements within this SP Single sheet and goes into detail on the decisions used within each section, the links of data, and formulas used where applicable.
4.3.2.8.1 Designation

The convention to dictate the specific stated preference combination is through the use of three different identification numbers; one for the origin FSA, one for the destination borough, and one for the stated preference scenario. To reiterate the total possible combinations available in the scope of the survey is:

- 205 origin FSA locations in the Greater Toronto Area – draws from Access sheet
- 19 boroughs in Montreal – draws from Egress sheet
- 6 stated preference scenarios – draws from Ngene sheet

To identify the FSA, corresponding city, a lookup function is used in Excel to output the corresponding names attached to the identification number. Based on the arrangement of the data in the previous sheets, a vertical lookup function was used. This vertical lookup function is a simple way to draw values from a data set without changing the original data set. This function is repeated to find the corresponding FSA name as well as the name of the destination borough.

In the following sections, a designation reference may be used to illustrate examples. The functional form of the designation is \([FSA\_Borough\_SP\ Scenario]\) where \([1\_1\_1]\) would designate a trip originated from the L1S FSA in Ajax going to Ahunstic-Cartierville in stated preference scenario 1.

4.3.2.8.2 Ngene Table

With the six (6) stated preference scenarios earlier defined and organized, it is necessary to draw the correct stated preference scenario from the six which matches with the designation in section 4.3.2.8.1. The resulting output of the Ngene table can be considered the main skeleton of the stated preference table that the other elements would attach to.

---

7 To output the corresponding city name for the FSA identification number 25, the formula =vlookup(FSA id value, Access sheet cells A2 to Q206, column 2), which would find the row with identification number 25 on the Access sheet and output the value on the second column of that row.
To copy the specific stated preference scenario from the Ngene sheet, a series of logic-based if-else statements are used to draw from the right scenario. The idea is to use the nesting feature of Excel formulas to incorporate a cascading series of logic-based queries\(^8\).

The use of if-else statements applies for most of the Ngene table with the exception for non-automobile access mode and non-automobile egress mode. In the pilot stage of the proposed project, it was discussed that non-motorized (walking) travel times over 60 minutes would not be logical. As non-motorized travel times were documented during the previous project stage, there is a record for walking trips from each origin/destination to and from the major arrival/departure stations. However, due to the large geographical spread, many of these non-motorized travel times are beyond the 60 minute barrier. It did not seem logical for an individual living in Ajax to walk seven hours to access Union Station. As five of the six stated preference scenarios in Ngene included either one non-motorized access or egress mode, an alternative procedure was required to make a more logical survey design.

The alternative used in the proposed project is to create a set of rules that would identify the appropriate situation to include or not include non-motorized local access/egress modes. The basic idea is to include non-motorized travel modes if the travel time is equal or less than 60 minutes. One disadvantage of this method is that a non-motorized travel time of 61 minutes would be omitted even though it is close to the 60 minute mark; however, it was decided that 60 minutes is an upper limit of an individual’s willingness to walk rather than the average acceptable time. When non-motorized local access/egress times were beyond 60 minutes, then a new travel mode is assigned. This new mode would be a randomized choice between the other local mode alternatives\(^9\).

---

\(^8\) One example to output the travel cost attribute level, the Excel formula used is =if(designation=1, output travel cost from SP scenario 1, if(designation=2, output travel cost from SP scenario 2, ... if(designation=6, output travel cost from SP scenario 6))).

\(^9\) The Excel formula used to randomize the new modal choice is =if(and(original travel mode is non-motorized, non-motorized travel time is over 60 minutes), randbetween(1 to 3 for local access or 1 to 4 for local egress), else stay with original value). To recognize the original local access/egress mode alternative, those values are stored under the access travel time in section 4.3.2.2 and under the egress travel time in section 4.3.2.3.
4.3.2.8.3 Access Travel Times

This element of the SP Single sheet is used to draw the specific local access travel times based on the origin location designation. The access travel time element displays the auto, transit, and walk travel times from the specified origin location to:

- Greyhound and Megabus departure station at Toronto Coach Terminal
- Via Rail departure station at Toronto Union Station
- Air Canada and Westjet departure station at Toronto Pearson International Airport
- High speed rail departure station at Toronto Union Station

The idea is to have all the relevant local access travel times available to potentially be matched with the relevant Ngene table element. For example, for the designation 40_5_4, the individual would access Pearson International Airport by transit (access mode = 2) with an access time of 150% baseline value. Referencing with the access travel time element, this local access trip would have a baseline travel time of 90 minutes, which translates to 135 minutes due to the 150% increase from baseline. While the automobile and walk travel times are not necessary for this scenario it may be used for other travel scenarios especially when the local access/egress mode is randomized due to a long non-motorized travel time previously discussed in section 4.3.2.2.

4.3.2.8.4 Destination

The destination element in the SP Single sheet is used to display the relevant intercity travel data from the Schedule sheet. For each of the intercity travel alternatives, the intercity travel distance, intercity travel time, daily departure frequency, and intercity travel costs are listed. The purpose is to use the values in this destination element as inputs into the Ngene table element to output the relevant information in the stated preference tables.

Currently the only intercity travel alternative is a trip starting from the Greater Toronto Area and ending at Montreal. The inclusion of this destination element, as well as the Destination and Schedules sheets are to set up for an expanded version where multiple intercity origin-destination pairs are included.
4.3.2.8.5 Egress Travel Times

Similar to access travel times, this element of the *SP Single* sheet is used to draw the specific local egress travel times based on the destination location designation. The egress travel time element displays the auto, transit, and walk travel times to the specified destination borough location from:

- Greyhound and Megabus arrival station at 1717 Rue Berri Street
- Via Rail arrival station at Toronto Central Station
- Air Canada and Westjet arrival station at Pierre Elliott Trudeau International Airport
- High speed rail arrival station at Toronto Central Station

The idea is to have all the relevant local access travel times available to potentially be matched with the relevant Ngene table element. For example, for the designation 121_17_2, the individual would arrive at Montreal Central Station by Via Rail. From the Ngene table, the individual would have originally traveled to Verdun by walking (access mode = 5) and arrive to the destination with the baseline travel time of 80 minutes. However, based on the previous rule imposed on walking trips over 60 minutes, an alternative egress mode was randomly selected. In this revised situation, the individual leaves Central Station by taxi (access mode = 3) and arrives at Verdun at 50% of the baseline travel time. As baseline travel time by automobile is 16 minutes, the individual total egress travel time in Montreal is 8 minutes.

4.3.2.8.6 Preliminary SP Table

The preliminary SP table element is the first instance where the all attributes for the stated preference survey are combined and output to something similar to the final SP table. While the arrangement of the attributes will be retained for the final table there is some attribute information that may be changed in later spreadsheets to help improve respondent understanding. One example would be the changing of time units to a uniform time format instead of a mixture of minutes and hours. Appendix I details how information in this table element is calculated from the amalgamation of all the previous elements in this *SP Single* sheet.
4.3.2.8.7 Itemized Row

In the pilot testing stage of the project, a potential problem was found in the procedures of preparing the stated preference tables, which would increase the amount of manual work required during the analysis portion of the project. The problem hinged on the alternative to randomize the local access and egress travel modes if it was originally stated as a non-motorized local access/egress mode with walking time exceeding 60 minutes. While the random generation of alternative local access/egress modes was logical, there were no additional procedures to document what each instance of randomization was. As a result, when analyzing the collected data, each stated preference table would need to be hand checked and updated with the correct values.

With the anticipated 500 survey responses and five out of six SP scenarios having the potential to require randomization, the additional work would be unnecessary and have the potential to incur errors. To mitigate this problem for the final survey, an intermediate step is required. The proposed solution is to create a large table that documents the information that will populate each of the 23,370 stated preference tables. In this method, if an access or egress mode is required to be randomized, the resulting randomization and corresponding local travel times will be stored in this intermediate database.

In the SP Single sheet, the itemized row is an element that translates the information from the Designation element, the Ngene table element, and Preliminary SP element into one row of information. Typically the values of the Ngene table were used unless it was an attribute that required time or cost information from the other elements in this SP Single sheet.

Another benefit of this intermediate step is the documentation of all relevant stated preference data for the purpose of modelling and data analysis. As each entry has a unique identification value (from 1 to 23,370) based on the designation value, it is possible to draw up the designation value from collected responses and pull up all relevant stated preference data.

4.3.2.9 SP Table

Drawing from the Itemized row element of the SP Single sheet, this SP Table sheet is the full database of all relevant stated preference information for each origin, destination, and SP scenario combination. The purpose of this SP Table sheet was to use the amalgamated data in the
later analysis stage as well as to use as an intermediate phase to store access and egress modes that required randomization.

The creation of \textit{SP Table} sheet is created through AppleScript automated software. The use of AppleScript was to go sequentially go through each of the 23,370 possible combinations and paste each entry from the \textit{SP Single} sheet into the \textit{SP Table} sheet. Afterwards, each FSA was given a unique identification number after being sorted. In this case, the sorting order (from smallest value to highest value) was by the destination ID, then by the origin ID, and finally the SP scenario ID.

A further discussion of the use of AppleScript is detailed in section 4.4. No additional calculations or data modification are required in the \textit{SP Table} sheet as the majority of the work is completed prior to this stage.

4.3.2.10 Lookup

The \textit{Lookup} sheet has a similar function as the \textit{SP Single} sheet where there the different aspects in the previous sheets are combined together to output a functional form of the final stated preference table for one origin-destination-scenario designation. As the name of the sheet suggests, the purpose is to lookup one specific designation and output the relevant cell values. Compared to the aggregating the data on a number of different sheets in the \textit{SP Single} sheet, the \textit{Lookup} sheet uses the intermediate \textit{SP Table} sheet as the primary database to output relevant information.

As previously explained, the \textit{SP Table} sheet is a database of all stated preference attributes for each of the 23,370 designations in the project scope. With each designation having a specific identification number, it is possible to output all relevant SP attributes in a vertical lookup if the correct identification reference is used. The functional form of this \textit{Lookup} sheet is to generate the unique identification number based on the designation inputs to output the relevant stated preference attributes. As the outputs from the \textit{SP Table} are still not in a presentable final form, another series of formulas are required to change the output data into contextual text.

The first function required is a formula to interpret the origin, destination, and SP scenario inputs into the relevant row identification number to match with the \textit{SP Table} sheet. In this scenario a vertical lookup (vlookup) function cannot be used as this function only matches one specific
value which would either be the origin, destination or SP scenario. However, the requirement is to find one specific value that that satisfies all three input values. A less intuitive, but more versatile, set of functions was used to remedy the limitation of the vlookup function. This set of functions can be referred to as the index-match function and uses binary logic to find the proper value. The index-match function is essentially two separate formulas where the outputs of the nested function become an input into the outer function. Appendix J describes this function in greater detail.

Obtaining the specific identification number correspondent with the input designation, a simplified version of the stated preference table can be generated by a vertical lookup of each element in that row. This new version of the stated preference table is similar to the one in the SP Single sheet but sources the information from a related, but different, data set. Figure 4-7 below compare between the two similar versions of the SP table.

From the feedback from the pilot test, there were suggestions to provide a summation of total travel time and total travel cost into the stated preference table. These summations were not previously to avoid these virtual attributes becoming the primary source of mode choice decisions. After the pilot stage, further discussion led to the decision to place total travel time and total travel cost into the SP tables realizing that it is an important aspect of decision and omission of it would induce survey fatigue when respondents attempt to manually find the total values during the survey.

One possible concern would be the magnification of errors if they were to occur due to the many processes used at this stage of the project. To minimize errors, all of these processes are automatic rather than reliant on human input after being designed. At each additional step in Excel, these formulas are checked multiple times for different situations for validity. The idea is
that if the formula passes the scrutiny of the initial verification process, then it should function with minimal error especially with the lack of external coding, such as VBA, required in the spreadsheet.

As the stated preference from the vertical lookup function does not fully show relevant information to the respondent, a revised version is required. This revised version is similar to the Preliminary SP Table element seen in the SP Single sheet and the outputs are also based on the Variable sheet where applicable. A revised method was devised as a simplified method to output relevant information without using multiple if statements. The content from the SP table on this sheet is used to vertical lookup the text string in the corresponding cells from the Variable sheet. This is applicable for the access mode, egress mode, transfers, seat choice, and trip information attributes.

Previously, local access and egress travel time was displayed as minutes and the intercity travel time was displayed as hours. A more unified method to display time is introduced at this stage to minimize confusion by the respondents. This new display format would display both local and intercity travel times as $X$ hours and $Y$ minutes. With travel times stored as decimal hour values, the method to place travel times into the proposed time format is detailed in Appendix K.

### 4.3.2.11 HTML

To display the stated preference table generated from the previous Lookup sheet on a web-based platform, it was decided that HTML coding would be the best method to output the information in a clean and professional manner. In this context, the tables would be stored as HTML code, which would be rendered by the respondent’s internet browser when the specific HTML code is called up. The alternative was to generate an image of the stated preference table and upload the graphic onto the web-based survey instrument. Some negatives associated with uploading the image is reduced quality when flattening vector-based text into a static image, potential issues with output depending on the scaling used to display content, and missed opportunity to place dynamic links and text within a graphics image versus HTML coding.

The purpose of the HTML sheet is to facilitate the conversion between the SP tables in Excel into an HTML format with minimal human interaction. Similar to previous Excel sheets translate information from one form to another form; this HTML sheet takes the information from the
Lookup sheet and places the appropriate data together in order to have a functional HTML output. A sample of the HTML sheet is available in Appendix L. The following paragraphs details how specific sections of HTML coding were incorporated into Excel.

Before the start of the table, introductory text is coded in, which informs the respondent of his or her progress through the six stated preference scenarios. This function was added from suggestions made at the pilot stage of the project. In addition, to verify origin and destination, the next text line states “From your home location in [origin city] to [destination borough]”. A hyperlink to a pop-up Google Map of the intercity trip is also included in this introductory section for reference. The purpose of this section is to provide some contextual information to the respondent regarding some aspects of the stated preference scenarios and to potentially reduce survey fatigue by informing the respondent of his or her progress.

Next, the stated preference table is placed into the HTML format. From testing, it was found that it was possible to create HTML code on separate Excel cells and have the proper outputs when copy and pasting the Excel HTML code into a text editor such as Microsoft’s native Notepad software. The majority of the HTML coding is the use of the table tags and corresponding table attributes. However, in addition to plain HTML, there are CSS coding elements included to simplify the coding of each individual cell. CSS stands for cascading style sheets and is primarily used to define how HTML elements are displayed (w3schools, 2013) by using references instead of reiterating display attributes every time attributes are changed.

<table>
<thead>
<tr>
<th>Access Method</th>
<th>Diving/Cancelling/Pasenger</th>
<th>Bus</th>
<th>Rail</th>
<th>Airplane</th>
<th>High Speed Rail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access Time</td>
<td>0h 0m</td>
<td>0h 15m</td>
<td>0h 24m</td>
<td>0h 0m</td>
<td>0h 0m</td>
</tr>
<tr>
<td>Egress Method</td>
<td>Pick Up (Cost = $0)</td>
<td>Rental (Cost = $50)</td>
<td>Rental (Cost = $50)</td>
<td>Pick Up (Cost = $0)</td>
<td></td>
</tr>
<tr>
<td>Egress Time</td>
<td>0h 0m</td>
<td>0h 15m</td>
<td>0h 30m</td>
<td>0h 0m</td>
<td>0h 0m</td>
</tr>
<tr>
<td>Percentage On Time</td>
<td>50%</td>
<td>60%</td>
<td>70%</td>
<td>70%</td>
<td>60%</td>
</tr>
<tr>
<td>Departures Per Day</td>
<td>6</td>
<td>4</td>
<td>21</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Number of Transfers</td>
<td>2 Transfers</td>
<td>Direct</td>
<td>1 Transfer</td>
<td>Direct</td>
<td></td>
</tr>
<tr>
<td>Seat Choice</td>
<td>Pre-booked Seating</td>
<td>Assigned Seating</td>
<td>Pre-booked Seating</td>
<td>First Come First Serve</td>
<td></td>
</tr>
<tr>
<td>Trip Information</td>
<td>Pre-Posted Schedule</td>
<td>Real-Time Schedule</td>
<td>Mobile Schedule</td>
<td>Real-Time Schedule</td>
<td></td>
</tr>
<tr>
<td>Main Travel Time</td>
<td>4h 24m</td>
<td>9h 42m</td>
<td>7h 36m</td>
<td>1h 12m</td>
<td>3h 0m</td>
</tr>
<tr>
<td>Travel Cost</td>
<td>$50</td>
<td>$27</td>
<td>$71</td>
<td>$225</td>
<td>$225</td>
</tr>
<tr>
<td>First Class/Business Class Cost</td>
<td>$80</td>
<td>$282</td>
<td>$304</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Travel Time</td>
<td>4h 24m</td>
<td>9h 54m</td>
<td>7h 54m</td>
<td>2h 5m</td>
<td>5h 12m</td>
</tr>
<tr>
<td>Total Travel Cost</td>
<td>$56</td>
<td>$32</td>
<td>$124</td>
<td>$275</td>
<td>$225</td>
</tr>
</tbody>
</table>
The initial CSS code is used to define the alignment of text and the output of the background color. As seen in the sample table above, the color for every other row is different to help respondents quickly compare information. In addition, the attribute description is aligned to the left while the information for each alternative is center-aligned. These simple display attributes help the respondent quickly look at and process information despite the fact that is a large amount of information for each SP scenario.

The basics of the HTML code are how each section (between the table row tag “tr”) is representative of one row of the table. To populate each row with relevant information, the HTML sheet references with the corresponding cells in the previous Lookup sheet.

4.3.2.12 TEXT

The TEXT sheet is similar to the HTML sheet in the way that the stated preference table in the Lookup sheet is used as the data reference of the attributes and levels. Rather than outputting into HTML code that is the equivalent of a table format, the TEXT sheet functions to output the stated preference table into a text based form. The sample text below corresponds to the same designation as the sample HTML stated preference table above.

Driving/Carpooling/Passenger
From your home in Toronto, a trip to Ville-Marie in Montreal costs $56 and is scheduled to take 4 hours and 30 minutes to reach your destination in Montreal. There is a 90% chance of the arriving on time in Montreal.

Bus
From your home in Toronto, you take transit, costing $3 to $6, taking 0 hours and 6 minutes to reach the Toronto Coach Terminal downtown.

From Toronto Coach Terminal, a one-way ticket costs $27 and is scheduled to take 9 hours and 45 minutes to reach the downtown bus terminal in Montreal. The bus service has 6 departures a day from Toronto Coach Terminal and will require 2 transfers to reach Montreal. There is a 80% chance of the bus arriving on time at the bus terminal in Montreal. When boarding, seats are pre-booked by passengers for passengers.

Arriving at the bus terminal in Montreal, you get picked up, taking 0 hours and 6 minutes to reach your final destination in Ville-Marie.

The overall door-to-door travel time is 9 hours and 60 minutes and the total one-way travel cost is $32.

Rail
From your home in Toronto, you walk or bike, with costs under $5, taking 0 hours and 18 minutes to reach Union Station.

From Union Station, a one-way ticket costs $71 and is scheduled to take 7 hours and 30 minutes to reach Central Station in Montreal. The Via Rail service has 4 departures day from Union Station direct to Montreal. There is a 70% chance of the train arriving on time at Central Station in Montreal. When boarding, seats are assigned at the terminal for passengers.
Arriving at Central Station in Montreal, you rent an automobile, costing around $50, taking 0 hours and 6 minutes to reach your final destination in Ville-Marie.

The overall door-to-door travel time is 7 hours and 60 minutes and the total one-way travel cost is $124.

Airplane
From your home in Toronto, you get dropped off, taking 0 hours and 24 minutes to reach Toronto Pearson International Airport.

From Pearson Airport, a one-way ticket costs $225 and is scheduled to take 1 hours and 15 minutes to reach Trudeau Airport in Montreal. The air service has 21 departures a day from Toronto Pearson Airport and will require 1 transfers to reach Montreal. There is a 70% chance of the airplane arriving on time at Trudeau Airport in Montreal. When boarding, seats are pre-booked by passengers for passengers.

Arriving at Trudeau Airport in Montreal, you rent an automobile, costing around $50, taking 0 hours and 30 minutes to reach your final destination in Ville-Marie.

The overall door-to-door travel time is 2 hours and 0 minutes and the total one-way travel cost is $275.

High Speed Rail
From your home in Toronto, you get dropped off, taking 0 hours and 24 minutes to reach Union Station.

From Union Station, a one-way ticket costs $225 and is scheduled to take 3 hours and 0 minutes to reach Central Station in Montreal. The high speed rail service has 30 departures a day from Union Station direct to Montreal. There is a 80% chance of the train arriving on time at Central Station in Montreal. When boarding, seats are available on a first come first serve basis for passengers.

Arriving at Central Station in Montreal, you get picked up, taking 0 hours and 0 minutes to reach your final destination in Ville-Marie.

The overall door-to-door travel time is 3 hours and 15 minutes and the total one-way travel cost is $225.

Due to the time limitation of survey distribution, this textual form of the stated preference table was not created for all the 23,370 different designation combinations; however, the functional form of this text-based output is complete and ready for future iterations of the survey.

4.3.3 Data Analysis Setup

The setup of the stated preference survey was designed to reduce the additional work required during the data analysis phase of the project; however, some additional work is required in order to organize the collected data into a functional form for data analysis and modelling. The purpose of this data analysis step is to define a set of procedures to turn the collected information from SurveyMonkey into a form suitable for modelling. While specific steps and decisions are based on the outputs from SurveyMonkey, the general procedure may be still applicable with data collected from other survey instruments.

Once again, Excel is the primary software used as a data management and modification tool. In the context of this project, the term modify does not suggest supplementing the collected data for something else. Rather, the term modify may be the expansion of data into another form (such as
the creation of dummy variables) or the interpretation of various data (such as inferring the number of local modes normally used by counting the respondents answers).

4.3.3.1 From SurveyMonkey

Two different outputs of collected data are available from SurveyMonkey. One is an output of the actual question choice text in the web-based survey instrument and the other is a numerical interpretation of the question choices. Both of these output types have the same functional output form where each potential question has an individual column of potential responses.

In the case where a respondent skipped a section due to the use of skip logic, the non-responded questions were left blank. For example, the survey build would advance the respondent to a generic stated preference survey if they indicated that they lived in a non-listed FSA origin. In the collected data, the first set of stated preference responses was left blank as SurveyMonkey automatically advanced to the generic set of stated preference responses. This method of storing respondent data benefits to normalize each respondent; however, condensing the data is required to prepare the data set for modelling. The procedure to condensing data will be discussed in the following sections.

One observation made during the pilot stage was the need to create an external document to list out how the numerical responses listed in the SurveyMonkey output corresponded with the text-based responses. Typically, the numerical choice set would always begin at one (1) and labeled according to how the text-based choice set was presented in this survey. Creating the list of corresponding values would save time in the next steps of data analysis setup.

An initial barrier to data analysis is the layout of the six (6) stated preference scenarios. The first potential problem is the horizontal layout of all the scenarios in one row, which is not the optimal arrangement. Instead, it may be better to separate each SP scenario into a different row. A second potential problem is the lack of relevant SP attribute information from the SurveyMonkey output. Because the SP table is externally linked, a procedure is required to recall the stated preference attribute information without incurring additional manual work.
4.3.3.2 Filter

As the proposed survey was passively distributed over the internet, it is not possible for the researcher to keep track of each respondent to know whether or not the survey was fully completed or completed to a point where it is still possible to use the respondent’s collected data for modelling. If an incomplete set of data was used for modelling purposes, it may result in an error during the computational stage. It was necessary to first filter out responses that do not meet the requirement for further analysis.

Currently there are two major filters used that have shown success at culling incomplete survey records. The first filter is the sixth stated preference scenario choice. This was chosen as the other five had to be completed in order for the sixth scenario to show up. Filtering for this last scenario would also mean that the resulting filtered data set would include all the stated preference choices and may be used if socioeconomic information was not required. The second filter was the household income question, which is the second last question of the entire survey. Once again, this question would not appear until all the previously required questions were completed. The completion of this household income question typically indicates that all previously required questions have been answered.

4.3.3.3 Combined List

This sheet functions as the primary step of outputting relevant information for use in data analysis and modelling. The main purpose of this Combined List sheet is to set up an Excel spreadsheet that dynamically collects the necessary information for each collected response. By putting in a valid respondent identification number, which is provided from SurveyMonkey, this Combined List sheet will draw from the necessary databases and modify the information in a way where the data is primed for further analysis. The following sections below detail some of the more intricate functions used.

As previously mentioned, all six (6) SP scenarios are horizontally recorded per respondent; however, it was beneficial for modelling purposes to separate the six scenarios into six separate entries. To split each collected response into six (6) responses, an empty Excel sheet was used to force the multiplication of the collected response list. Through the use of vertical lookup table, it
was possible to match the original list of responses with a new list that increased incrementally every six (6) rows.

The majority of the data in the Combined List sheet are simple vertical lookups to reference a specific column on either the numerical or actual text outputs from SurveyMonkey. While it was resource intensive, only one entry has to be manually entered as all subsequent entries will be based on the formulas in that first row.

While most of the columns were simple vertical lookups, there were instances where more advance functions were required. In SurveyMonkey, FSA designations were listed based on the respondent’s origin city. In SurveyMonkey’s output, the respondent’s FSA choice was placed into separate columns dependent on the chosen origin city. An automatic system to amalgamate the different columns into one single column was required. In Excel, a series of functions was required to output the first instance of a response in a range of specified columns. This was done by using a variation of the index-match function described earlier in section 4.3.2.10 and the indirect function.

- The indirect function returns the reference specified by a text string and is typically used when the reference to a cell is required to change without changing the formula itself (Microsoft, 2013). In this project, the indirect function is used to reference a string of text which is created to reference the spreadsheet and relevant row pertaining to the origin FSA questions but also referencing with the row value of the respondent ID from the original SurveyMonkey data. In this case the indirect function has to be used as the row value of respondent ID is dependent on the position on the original data set and not the position of the ID on this sheet.

- The index-match function searches across the specified row from the original data set and finds the first instance where there is text. When there is text, then that text is displayed on the Combined List sheet and the function stops.

As previously noted, the outputs from SurveyMonkey do not include the stated preference attributes. Therefore, it was necessary to have a system to draw from the stated preference information used when constructing the HTML tables. In this case, two sets of data were included in the Combined List sheet; the attributes of the chosen stated preference intercity
modal alternative and the attributes of all possible alternatives for the given SP designation. These attributes of the chosen mode is repeated in the full listing; however it simplifies the analysis process by separately listing the attributes of the chosen mode. A facsimile of the *SP Table* sheet used in the SP table generation workbook is located in this data analysis workbook.

- A match function is used to find the row from the SP Table sheet specific to the SP designation, which is obtained by looking at the origin FSA text, SP destination choice, and SP scenario number.
- The indirect function is used to reference the obtained row identifier with the proper mirror outputs from the *SP Table* sheet.

### 4.3.3.4 Dummy List

The *Dummy List* sheet is a variation of the *Combined List*, which separates some of the collected information into a series of dummy values. When an attribute is not considered continuous the use of non-dummy values may influence the outcome of the resulting model. For example, travel time is considered continuous and enters into the model; however, seating choice is designated by an integer value between one (1) and three (3). In this case, keeping the attributes as this ordered list of values implies that the order of how it is listed has importance, where first-come seating (designated as 3) is higher than assigned seating (designated as 2). To eliminate this bias, creating dummy variables would separate the attribute into three attributes designated by binary values.

To create dummy variables, a series of if-statements are used and refers back to the corresponding attributes on the numerical or actual text sheets. As the stated preference attributes are generated using the indirect function, a series of additional dummy variable columns are added alongside the non-dummy variable stated preference attributes.

### 4.4 AppleScript

In the current setup, only one designation (out of the possible 23,370) is displayed in both the *HTML* and *TEXT* sheets. The process of copy and pasting the HTML code for each designation is a feasible but resource intensive task that could benefit from automation. Based on the researcher’s use of Apple computing hardware, the scripting software AppleScript was used to create an automatic process of generating the stated preference tables in the HTML format.
AppleScript is a versatile scripting language built into the OS X software. The scripting language is possible to control scriptable applications as well as many parts of the operating system. A scriptable application is one that can make its operations and data available in response to AppleScript coding (Apple, 2007). Several uses of this include creating shortcuts, automating repetitive tasks, or creating custom applications with the general purpose of saving time (Apple, 2013). The following steps are a list of repetitive tasks required in generating one stated preference table in HTML format:

1. Select 1 of 19 destination values in Excel
2. Copy the contents of either the HTML or TEXT sheet
3. Paste into a text editor
4. Save the file with the proper designation under the HTML file format
5. Increment the designation
6. Repeat steps 2 to 6

From initial trials, it was found that the coding used was possible to generate the required stated preference tables in HTML format; however, continuous use of the custom AppleScript code would cause memory leaks and cause errors in the coding. Using trial and error, it was found that AppleScript can operate continuously in the generation of all combinations of origin-destination-SP scenario for one destination value, which are 1,230 iterations. When each destination was completed, the computer would be power cycled to clear any possible memory leakages.

As the AppleScript coding for the stated preference tables and text are similar, only the coding for the generation of SP tables are shown. The benefit in using the AppleScript language to automate this task is the close similarity between spoken language and the AppleScript language. Seen in Appendix M, the text is based on basic text and not cryptic functions. In addition, a series of nested functions were required; however, it was easy to define different function nests without using too much coding syntax. Finally, the main benefit of AppleScript is the ability to call basic functions such as dictating certain keystrokes or wait times.

The one limitation of using AppleScript is that the computer cannot be used when the code is running as it would interfere with the coded functions. However, despite this limitation and the limitation with the memory leaking, the use of AppleScript was a quick and simple way to automate this largely repetitive process while inducing essentially no error.
Chapter 5  
Web Survey Design

A web-based survey instrument is necessary to implement the proposed survey design. This web-based survey instrument can either be designed from scratch or utilize existing services provided by third-party companies. This survey instrument should follow the project objectives previously stated in section 1.1. To reiterate, the objective is to utilize an efficient web-based survey instrument that does not rely on traditional survey methods and can be easily modified and adapted for other research purposes. The following sections detail the alternatives explored for the web survey instrument as well as elaborate on the benefits and limitations of the chosen alternative.

The proposed web-based survey design will be referred to as the survey build. Compared to the survey design, which are the design choices made in deciding the structure and questions within the survey; the survey build are the choices made when applying the survey design into a web-based survey instrument.

5.1 Alternatives

Three prime alternatives were explored in the process of designing the web-based survey; a novel php-based design, through the third-party service Qualtrics, and through the third-party service SurveyMonkey. While the alternatives explored were not an exhaustive list of potential survey instruments, these alternatives represented the ones with the highest utility towards implementing a complex survey design. Utility in this case is a balance between the resources (cost and time) required to create the survey and the features available with a specific service.

Aside from the coding from scratch alternative, other considered third-party survey services were considered from different sources, including general web queries and academic literature. Third-party services used by other companies and academic institutions were preferred. In addition services with the option for survey panel integration were also considered as an asset. Table 5-1 below is a summary of some third-party services that were considered based on reference from Marra’s conference proceedings (Marra & Bogue, 2006) but were not investigated into at depth given time limitations.
<table>
<thead>
<tr>
<th>Web-based Survey Service</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualtrics</td>
<td><a href="http://www.qualtrics.com">http://www.qualtrics.com</a></td>
</tr>
<tr>
<td>SurveyMonkey</td>
<td><a href="http://www.surveymonkey.net">http://www.surveymonkey.net</a></td>
</tr>
<tr>
<td>SurveyGizmo</td>
<td><a href="http://www.surveygizmo.com">http://www.surveygizmo.com</a></td>
</tr>
<tr>
<td>Sawtooth Software</td>
<td><a href="http://www.sawtoothsoftware.com">http://www.sawtoothsoftware.com</a></td>
</tr>
<tr>
<td>LimeSurvey</td>
<td><a href="http://www.limesurvey.org">http://www.limesurvey.org</a></td>
</tr>
<tr>
<td>Vovici</td>
<td><a href="http://www.vovici.com">http://www.vovici.com</a></td>
</tr>
<tr>
<td>Nooro Online Research</td>
<td><a href="https://www.nooro.com">https://www.nooro.com</a></td>
</tr>
<tr>
<td>Allegiance</td>
<td><a href="http://www.allegiance.com">http://www.allegiance.com</a></td>
</tr>
<tr>
<td>SurveyWriter</td>
<td><a href="http://www.surveywriter.com">http://www.surveywriter.com</a></td>
</tr>
<tr>
<td>SurveySystem</td>
<td><a href="http://www.surveysystem.com">http://www.surveysystem.com</a></td>
</tr>
</tbody>
</table>

Table 5-1 – List of Web-Based Survey Services Considered

Discussion regarding a web-based survey instrument will utilize the terms “front-end” and “back-end” systems, which are terms within computer science to discern between the processes used when interacting with a user. Within the scope of a survey instrument, back-end systems are the components that process information and interact with databases. In contrast, front-end systems interface between the user and back-end systems by collecting inputs from the user and concurrently presenting to the user relevant information processed from back-end systems. Recognition between these two types of systems is important to assess the advantages and disadvantages of the various survey instrument alternatives.

5.1.1 Coding from scratch

This alternative is described as creating the survey instrument using custom coding. In general, this method of creating a survey instrument allows for the most complexity to be incorporated into the survey; however, may not incorporate a redundant data storage system or be as visually appealing to the respondent.

The main benefit of this alternative is full creative and technical control of the survey instrument. Limitations in the type of data processing are only limited by the expertise of the programmer creating the survey. While the front-end interface and back-end database systems can be sourced from existing projects, there is the freedom to manipulate the code to adapt to a required function. For example, it is possible to ask the respondent to input an experienced travel time (front-end), multiply this inputted travel time with a specified coefficient (back-end), then display this modified travel time within a stated preference table.
A variety of programming languages may be required to create a custom survey. Due to the large number of programming languages, the following list is not an exhaustive list nor is it an exactly list of what is required.

- PHP – as the fundamental backbone for collecting, processing, and presenting data
- Java – an alternative function to PHP that could be used synchronously
- HTML – basic visual component of the front-end system
- CSS – advanced visual coding to improve the survey experience

One of the obvious limitations to a custom coded web survey instrument is if the survey designer does not have the necessary background in programming. If this is the case, then additional resources may be required to either have the survey coded by a third-party or to hire an individual with programming experience. If an external source is required to provide the coding, any further changes or improvements to the survey code may not completed without contacting the external source, creating further delays and increasing development resources.

Another potential disadvantage to custom coded web survey instruments is the lack of adaptability. Depending on the scope of the project, it is possible that the coding is rigidly tied to the proposed function of the survey. A survey may be coded to have specific functions that are non-essential to another survey scope but is integrally tied to the main coding and cannot be removed. Much like writing styles, two individuals may have distinct coding styles and modifying another individual’s code could present an unnecessary barrier when there are other alternatives available. While this disadvantage would not deter from the function of the proposed survey design, it may be a disadvantage to researchers who create multiple surveys on similar topics.

In the early stages of this proposed intercity survey design, an adaptation of a fellow research colleague’s survey was considered; however, as the scope of this intercity survey changed, the adaptation of the existing code was becoming more of a barrier than an asset. In addition, the additional programming resource was not available in the scope of this project.
5.1.2 Web-based Service

As previously stated, there are multitudes of web-based survey services available on the internet. With so many choices, the determining factors used within the scope of this project were the cost of the service and the availability of survey tools. Availability of survey tool is defined as the number of tools at the disposal of the survey designer to design, modify, implement, and analyze his or her survey design. These tools can range from the type of questions available for the survey designer, the ability to change color schemes, to how the web-survey can be distributed. While there is no specific metric or guide to quantify a web-based service, the service was first judged on the available tools to the designer followed by the cost of service.

Another attribute considered was the existing client base of the service provider. Having a large and well-known client base was an indication that the survey service has seen previous success with both commercial and research based projects. From the web-based survey services explored, Qualtrics and Survey Monkey were chosen to be explored further based on their function, price, and existing client base. The following sections detail both of the explored services in terms of building the survey, distributing the survey, and using survey data for analysis.

5.1.2.1 Qualtrics

Qualtrics is an online service provider that primarily focuses on the creation and management of customizable surveys. The basis of these types of online web survey instrument services is the emphasis on a user to create a survey using the tools that the service provider has created for the specific company. Referencing to the different types of systems, Qualtrics has a fully functional front-end and back-end system as soon as a user signs up for the service. The only thing that a survey designer has to do is to apply his or her survey design into Qualtrics. That being said, the complexity of the survey design is limited to the programming of the front and back end systems as dictated by the programmers working at Qualtrics.

In regards to building the survey, Qualtrics offers a large range of question types. According to Qualtrics, there are over 100 question types offered including the typical multiple choice and text entry options. Within the scope of the proposed intercity survey, some of the unique question types offered by Qualtrics are:
- Time-limited questions to push the respondent to make a quick choice
- Drill down questions that dynamically change based on previous choice within one question rather than having to answer multiple pages of questions
- Ranking of choice based on dragging and dropping text or images
- Heat map of an image reveals an aggregated result of a location on a map or seating choice
- Ability to make a choice on a graphic, which may be useful for map-based navigation

The large number of non-typical question types is beneficial to potentially provide novel ways to collect information from respondents. Rather than reading text and selecting a check box, the more graphical question types in Qualtrics may make going through a survey less repetitive. However, it is possible that introducing too many different question types may confuse a respondent unless specific instructions are given. This confusion could lead to inaccurate results or increase survey fatigue.

Beyond the number of question types available to the survey designer, a large choice of customization both in terms of presentation and incorporating audio and visual components are possible within Qualtrics. It is possible to provide respondents the possibility to download files, which may be helpful to further inform the respondent based on their specific response. If incorporating some sort of incentive, such as discount coupons on travel tickets, respondent-specific incentives can be delivered based on Qualtrics’ design.

Another important aspect in Qualtrics’ design is the use of logic-based questions. One of these is to carry forward choices, which is the ability to store choices from previous questions and to display what was, or was not, selected later in the survey. Another logic-based tool is the ability to skip questions or blocks of questions depending on a respondent’s choices. There are multiple application of logic with the more advanced logic flows utilizing a “then if” branch, which controls the respondent’s survey flow and has the ability to create multiple termination points. The advantage in having logic-based flow is to reduce a respondent’s potential for survey fatigue by only showing questions relevant to his or her choices. For example, if the respondent has not made an intercity trip within the past year, then the block of questions regarding intercity revealed preference does not need to be asked as the respondent did not qualify the necessary condition.
Compared to Survey Monkey and other popular web-based survey services, Qualtrics has the largest amount of customization available for a survey designer. While not all the question choices or system logic is relevant to the proposed survey design, the breadth of choices available is helpful in the survey creation process as well as to maximize the attention span of a respondent.

With the stated preference design being a key component in the proposed survey, it was important that the chosen survey instrument can incorporate the proposed stated preference questions with relative ease. One of the main reasons why Qualtrics was explored in-depth was its advertised conjoint analysis options. When developing the pilot survey, it was found that the back-end system was not accessible by the survey designer. As a result, it was not possible to have background calculations such as multiplying a baseline value with a back-end stored coefficient table. While the sales team and technical service support was always helpful to look for possible solutions, the complexity of the proposed SP survey was a major limitation despite the other advantages in Qualtric’s survey design.

In regards to the distribution of a Qualtrics survey, there are a number of distribution methods including but not limited to; email, web link, survey panel, and local servers. When links are sent out into the public, it is also possible to track each response in real time to assess completion status as well as view an individual’s pass invitations. To increase response rate, scheduled reminder messages may be sent out at a defined time and thank you messages can be programed to be sent upon survey completion.

Qualtrics also has a set of basic data analysis tools that may be beneficial to quickly compare between different demographic groups or create statistics tables without going into the raw responses. There are numerous graph types as well as cross tabulation results, which are helpful to sort a large number of data. Additionally, a simple conjoint analysis tool is available; however, the proposed stated preference design is more computationally advance to the analysis tool available on Qualtrics.

Overall Qualtrics provides a feature-rich package of survey design, distribution, and analysis. The area that Qualtrics excels in would be the survey design because of the large number of question types available in addition to the advance implementation of survey flow and logic-based survey design. On the distribution side, Qualtrics has many choices available but does not
introduce too many additional features when comparing with competing services. The data analysis feature in Qualtrics is useful to have a quick grasp at demographics; however, advance data modelling is not featured. In the trial period of using Qualtrics, the main limitation is the cost of the service. Compared with competing services, Qualtrics does cost more and offer less in terms of the number of allowable survey respondents. If Qualtrics had a method of back-end calculations, then the cost may have been more justified; however, the limitations experienced with creating the stated preference questions in Qualtrics put it on par with more cost-competitive services.

5.1.2.2 SurveyMonkey

SurveyMonkey is one of the most used web-based survey services currently available. Similar to Qualtrics, SurveyMonkey is an online service provider that primarily focuses on the creation and management of customizable surveys. SurveyMonkey uses a functional back-end system and integrates with a satisfactory front-end system. The features available in SurveyMonkey allow for a survey designer to integrate his or her design within the established front and back-end systems; however, there are limitations that hinder the survey design process.

One of the reasons why SurveyMonkey has a large user base is the inclusion of a free basic level account option, which limits a user to create a survey with less than 10 questions and collect as many as 100 responses per survey. The availability of this basic level account is a benefit for non-power users to create simple surveys while promoting the SurveyMonkey brand. Going into the paid level accounts, the prices are reasonable given the features that are included in both the “select” and “gold” account options. For this project, a gold level account was used to create the proposed intercity survey and the following discussion relates to features and options available at that account level.

The survey building process is fairly straightforward with a number of question types available to the survey designer. According to SurveyMonkey, there are 15 question types, including the basic multiple choices, matrix of choices, and text entry. Compared to the question types offered by Qualtrics, the SurveyMonkey question types are geared towards a text based interface. While there is added utility in incorporating visual media and interaction with visual media within the survey design, the lack of these extra question types did not hinder the survey building process. Additionally, visual media based question may be useful for collection non-essential data, but
may hinder the collection of essential data such as location information. For example, while a heat map of respondent’s home location is visually appealing to put into a report or presentation, there may be errors incurred when cross referencing with access time. Comparatively, there is less potential for error if the user chooses a multiple choice option of their home FSA address.

Basic customization options are available in SurveyMonkey and can be easily accessed and changed if required. The look and feel of the survey, such as color and font, are reflected within the design interface. In addition, other options, such as the inclusion of a progress bar are easily accessible within the survey builder. One limitation experienced during the survey building phase is the difficulty to create copies of surveys. There is no method to import/export designed surveys to be stored on the user end for exchange or backup. Additionally, the only method to create a new version of the survey is to copy an existing survey. As there are no systems to download surveys, previous survey versions are stored in the account and arranged chronologically based on last modified date. While it may not appear to be a limitation for single account holders, conflicts may be caused when multiple users building multiple surveys on a single account.

Similar to Qualtrics, logic-based question flows are available in SurveyMonkey. While these logic-based flows are available the implementation in SurveyMonkey is more rudimentary when compared with the logic features in Qualtrics. In SurveyMonkey, logic-based flows are only available on multiple choice questions. For example, one question in the proposed survey inquires about the frequency of the respondent’s intercity travel by using a matrix of choices; however, the response on this question cannot be used as a logical skip if the respondent has not made an intercity trip within the last year and a similar question has to be asked additionally for the logical skip to occur. In addition, SurveyMonkey does not have a visual representation of the logical skips used within the survey design. While the lack of visual representation does not impact in the function of the survey, the survey designer is required to check each question to verify the skip logic used.

With the stated preference design being a key component in the proposed survey, it was important that the chosen survey instrument can incorporate the proposed stated preference questions with relative ease. SurveyMonkey did not advertise that it was possible to create a survey for conjoint analysis; however, when Qualtrics did advertise conjoint analysis as a
feature, the proposed stated preference design had too much computational complexity. Similar to the pilot development in Qualtrics, development of a second pilot survey also encountered limitation due to lack of access to back-end database calculations. While a method was eventually found to incorporate the stated preference design into SurveyMonkey, this method should also be applicable to work with Qualtrics. Detailed methodology into building the stated preference survey in SurveyMonkey is discussed in section 5.3.

In regards to distribution of a SurveyMonkey survey, response can be collected by web link, email, website, and Facebook. Similar to Qualtrics, when links are sent out into the public, it is also possible to track each response in real time to assess completion status as well as view an individual’s pass invitations. One of the main benefits of SurveyMonkey’s implementation of survey distribution is the management of response collectors. A number of collectors can be issued for one survey and recorded as a specific collect identification. The benefit of tracking collectors is the potential to segment demographics based on the collector. Segmentation may be useful given this non-traditional approach at defining sample frame.

One negative aspect of SurveyMonkey is the name of the service. In attempting to promote an academic study, the implication of the word “monkey” within the title and branding was predicted to have a negative effect on potential respondents. Potential respondents may not take the study seriously given connotations with the monkey term. Upon further use, a large number of commercial companies also use the SurveyMonkey service and retain the SurveyMonkey branding. While non-academic in nature, the widespread use of the SurveyMonkey service by companies such as Audi and Samsung (SurveyMonkey, 2013) do alleviate negative connotation towards the name.

Like Qualtrics, SurveyMonkey also has a set of basic data analysis tools that may be beneficial to quickly compare between different demographic groups or create statistics tables without going into the raw responses. Filters can be defined and cross tabulation can be done within SurveyMonkey; however, given the complexity of the proposed survey it was more convenient to download the raw data and analyze in external spreadsheet-based software.
5.2 Survey Instrument Selection

SurveyMonkey was selected as the web-based survey instrument for the building and distribution of the proposed survey. SurveyMonkey was chosen as it offered the survey designer the best ratio between flexibility and cost. While SurveyMonkey lacks the assortment of question types offered by its competitor Qualtrics, the question type set within SurveyMonkey was deemed satisfactory to build the proposed intercity survey. The visual-rich question types in Qualtrics could be beneficial but were not essential to the function of collecting data. In terms of cost, SurveyMonkey’s pricing structure offered more within this scope of this proposed project and other projects than Qualtrics. Compared to the custom coded survey option, the lack of coding resources and the limited time to complete the project were the deciding factors to choose an existing web-based survey service.

SurveyMonkey has provided an intuitive survey instrument that has the ability to support an intricate survey build, such as the proposed stated preference/revealed preference intercity survey. Despite its strengths, there are also weaknesses that are inherent in the survey build and potentially hinder the survey design. The sections below expand on the benefits and limitations of SurveyMonkey that were not discussed in depth previously.

5.2.1 Benefits

The front-end system in SurveyMonkey can be defined as the user interface for both the survey designer as well as respondents. From the survey designer point of view, the usability of the front-end system is critical to building and managing web-based surveys. The survey creation interface outlines the important elements, such as where to add questions or pages, while keeping extraneous information to the sides. In addition, having contextual elements, like adding/editing a question, as a pop-up within the webpage maintains continuity instead of bringing the designer to a separate page each time a button is pressed. The design elements incorporated by the SurveyMonkey team lend help to an individual creating a survey using SurveyMonkey.

In addition, organizational front-end system tools available to a survey designer are beneficial to building the survey. SurveyMonkey has included tools to organize multiple surveys for the account holder as well as tools to organize questions within a survey. For example, it is possible to organize different questions within pages. A page can display a number of questions and
typically only advances in the case of a new set of questions or when a question with a logical branch deviates based on user input. The SurveyMonkey interface can also view the survey by all pages and surveys or a single page, allowing the survey designer to have a general view or to focus in on a specific section.

From the respondent’s point of view, a well-organized survey with a logical progression of questions reduces the number of possible ambiguities as well as reduces the perceived time spent completing the survey. Specific grouping of questions within a page is intended to create a sense of unity in the questions without having the survey designer explicitly state that a sub-set of questions are related. A dynamic progress bar is also an available visual element that can be included within the survey to show the respondent an approximate completion rate.

Based on the choices in terms of font, font size, and color scheme, SurveyMonkey can potentially offer a very clean and minimalistic user interface for the survey respondent. When going through the survey, a respondent is only provided with the necessary information; the question, possible responses, the advance button, and a survey exit button. Having a lack of distracting information aims to reduce survey fatigue and possible confusion to the respondent. For example, for questions that require a reply, a non-reply from the respondent would prompt the display of red text indicating questions that require a response. These small visual cues are helpful to progress through the survey flow. The ability to specify various design elements also helps to improve the branding of the survey to give a unified identity between the survey and survey group. The SurveyMonkey front-end system allows for efficient question/page layout to be used in conjunction with visual indicators to reduce the burden for both the survey designer and respondents.

In traditional pen and paper surveys, if a set of questions is not relevant for the respondent, it was still required for either an individual marking down responses or the respondent to read which section they would or would not have to go to, based on a question response. This unnecessary work has the potential to disrupt survey flow or collect incorrect/unnecessary information. Moving to an electronic-based survey platform, it is possible to present information dynamically during data collection.

As previously explained, the application of basic logic statements in SurveyMonkey is a benefit in terms of creating efficient survey flow. While the implementation of logic-based flow is not as
robust as Qualtrics, it is satisfactory in adapting the survey design. Even though logic-based question skipping is only available for single-answer multiple choice question types, the SurveyMonkey interface is intuitive to selecting a specific question branch. One example of the application of logic-based questioning is to inquire about the respondent’s forward sortation area (FSA). Currently, there are approximately 205 FSA located within the Greater Toronto Area. While a simple drop-down list of all possible FSA choices could be used, it would cause some burden for a respondent to navigate through all the choices. Instead, it is possible to narrow down a smaller list of FSAs based on the respondent’s city using logic-based question skipping.

Another application of logic-based question skipping is inquiring about a respondent’s most recent Quebec City – Windsor Corridor (QWC) trip within the past year. If the respondent indicated that they have not made a QWC trip within the past year, it would only cause increased burden to ask the respondent about extraneous information. As a result, SurveyMonkey will skip the entire intercity revealed preference set of questions. In both case, the use of logic-based question skipping is to exploit the ability of the web-based survey instrument to dynamically present only relevant information to the respondent than to burden him or her with extraneous information.

The ability to utilize HTML coding in some form is a benefit of SurveyMonkey compared to more basic web-based survey services. The use of HTML coding varies depending on intended purpose; however, the inclusion of some basic coding expands the creative ability of a survey designer to incorporate elements that may not be already included. One example would be the use of HTML iframes to externally link the stated preference tables within SurveyMonkey. This is explained in depth in section 5.3.

In the proposed survey, there are no defined lists of individuals to draw responses from (further explained in section 6.2). With the entire GTA as both the population frame and sample frame, respondents are contacted via various groups that exist online such as Facebook, Reddit, email listservs, etc. It is predicted that these groups have various socioeconomic structures which may influence modal choice and a system is required to discern between the different groups. While it is possible to directly ask the respondent the source of the link, there is added benefit to create multiple entry points into the same survey and distribute those entry points based on the different
online groups. In addition, these links can be customized to have phrases instead of random characters, increasing the legitimacy of the survey.

5.2.2 Limitations

Despite all the benefits of SurveyMonkey, a number of limitations were encountered during the web-based survey build. While most of these limitations were not detrimental to the application of the survey design into the web-based survey instrument, non-intuitive methods to circumvent these limitations had to be created.

The ability to organize questions and pages without much hassle is a benefit; however, the lack of an intuitive ability to copy and move questions/pages is a disadvantage especially when creating long surveys with multiple logic branches. While there is the ability to copy and move, a drag and drop system would be much easier to visually indicate what is being moved. Currently, if a question/page is selected to be moved or copied, SurveyMonkey will store the selected question and prompt for a possible location to move the selected item. If the designer forgot which question was selected or wanted to double check, it is not possible without either cancelling the action or committing to the move to find out. While this is a minor inconvenience, errors may occur when organizing a long or multi-branched (non-linear) survey.

As previously explained, SurveyMonkey can organize a number of questions within a defined page. If a logic-based function is used for a question, then the resulting branches of questions have to be placed on a separate page. However, there is no method to group a number of pages together. One way to get around this is to name each page with a category. For example, the four individual sections of the proposed intercity survey are listed from parts A to D. While this is a rudimentary but effective way to identify each page, it would not be possible to move entire groups of pages together or view only the questions in one group of pages.

The previous two limitations deal with the organizational tools available to the survey designer. The development of a survey flow interface, which details any potential logical skips and question hierarchies, would be beneficial given complex surveys with multiple logic branches as well as distinct question groupings. Implementation of this indication of survey flow is not impossible given the option being available on Qualtrics’ service. While the lack of this function
is not detrimental to the success of the survey instrument, it could be useful for design and interpretation of a complex survey.

The customizable elements and the main interface are two beneficial features of SurveyMonkey that can be defined as front-end system elements. Despite having a very flexible and customizable front-end, the back-end system is closed off to the survey designer. An ideal back-end system would support an open system for survey designers to create additional functions to incorporate into the web-based survey instrument. However, given the cost of service for SurveyMonkey and the intended use of the service, creating an additional platform to support a few additional functions by specific research groups may not be the best intended use of resources for SurveyMonkey.

It is possible to upload images into SurveyMonkey and have the image appear as part of a descriptive text; however, there are no storage options for non-image file types. Rather than creating 20,000 different questions in SurveyMonkey for the stated preference tables, HTML iframes could be used to reference the SP tables based on previously answered questions (further explained in section 5.3). However, due to a lack of storage options within SurveyMonkey, these HTML table files had to be stored on another server. If that external server was shut down for maintenance or experienced technical difficulties, then the SP tables would not show up in SurveyMonkey. There are existing server space at SurveyMonkey to store files such as photos, it was a limitation to have to search for an external host to store non-image files and potentially experience technical difficulties to access the SP tables rather than having all the files under one host server.

Despite the limitations listed, SurveyMonkey is one of the most capable web-based survey services available. The options and customization available to the survey designer is capable of translating the proposed survey design into a web-based survey.

5.3 Adaptation of External Coding

The stated preference section of the proposed design is the main focus of this intercity mode choice survey. To reiterate, the purpose of a stated preference set of questions is to gauge how various attributes may affect mode choice. By presenting a respondent with different scenarios with varying attribute levels, it may be possible to impute mode choice by finding the attributes
that have the greatest impact on an individual’s utility. To collect relevant data for analysis, it is important to present relevant information to the respondent. In the proposed project, one of the objectives is to collect geographically disaggregate information to assess how access and egress may impact intercity mode choice. As a result, the scope of the stated preference portion is computationally complex, which presents a challenge when using a web-based service such as SurveyMonkey.

In early developments of the proposed intercity mode choice survey, access was disaggregated to five (5) different locations around the Greater Toronto Area, which were the approximate centers of major urban areas such as; downtown Toronto, Mississauga Square One, etc. Additionally, this early survey design had one intended destination, McGill University in Montreal, which was assumed to be the most known location. Under these factors, it was not exceedingly prohibitive to create 30 individual tables for each OD pair.

In the final survey design, emphasis was placed on geographical disaggregation. The number of origins expanded to 205 FSAs and the number of Montreal destination expanded to 19 boroughs. In total there were 23,370 combinations of OD pairs and SP scenarios. The first alternative was to create 23,370 individual questions within SurveyMonkey. Using this method, the tables would be typed out as HTML code and would be placed as descriptive text. A large series of skip logic would be used to advance respondents from their selected FSA designation to the relevant six (6) stated preference scenario tables. While it may be tedious, to create the large number of questions once would still be within the realm of possibility given a full week of constant work in SurveyMonkey. In this case, the main disadvantage would be to have a change that affects the entire set of stated preference questions, as each 23,370 questions would have to be modified. In addition, the large number of stated preference questions would make navigation of the entire survey difficult to the survey designer. A more efficient method to incorporate a large amount of data within the limitations of the SurveyMonkey environment is required.

To circumvent this abundance of work required incorporating the numerous permutations and combinations of origin/destination pairs, the use of HTML coding and question piping was used. In this context, HTML coding is defined as the ability of SurveyMonkey to display information to the respondent based on HTML coding rather than through pre-set design elements. Question piping is defined as the function of SurveyMonkey to store a previously answered response and
output that response later in the survey. The use of question piping is indicated by the typing out the question number in between brackets. While question piping may not be an exclusive feature to SurveyMonkey, the details provided in this paper are related to SurveyMonkey’s use of question piping.

The general idea of presenting the respondent with specific origin/destination based stated preference tables is to store all possible tables somewhere on the internet and then query a specific file based on a respondent’s choice. As previously described in section 5.2.2, SurveyMonkey does not have an advanced file hosting interface. Consequently, these HTML files were stored on a server space hosted by a fellow colleague under the address logaday.org. When storing the HTML on the logaday.org server, the file hierarchy is indicated by Figure 5-1 below.

![Figure 5-1 – Hierarchy of Information Stored on Logaday Server](image)

Under this hierarchy, it is possible to arrange for different origin/destination pairs without the need to create long filenames for the HTML tables. The HTML naming convention is based on the FSA address and the SP scenario. While the addition of the destination (borough name in Montreal), a long complex name such as Rivière-des-Prairies—Pointe-aux-Trembles may have the possibility of generating errors based on the use of non-Roman characters. Placing the HTML tables based on folders named after the destination city and borough was assessed to reduce errors and increase organization. In addition, if the scope of the proposed survey was to be expanded to having multiple cities as possible intercity travel destinations, then this approach
at organizing the HTML tables would easily facilitate the expansion through the addition of more folders within the logaday.org server.

There are two main uses of HTML code in SurveyMonkey when presenting a respondent with relevant stated preference tables; creating an iframe to show an external HTML source and matching the right HTML table based on the respondent’s previous responses. An iframe is defined as an inline frame that is used to embed another document within the current HTML document (w3schools, 2013), much like how a picture frame can display something external to the context of the environment where the picture frame is hung. In addition, various attributes can be modified on the iframe using HTML tags. In this case, attributes were selected to make the iframe integrate into SurveyMonkey seamlessly to give the respondent the perception that the stated preference tables are being generated in SurveyMonkey rather than through the external logaday.org site. The pixel width and height of the iframe was modified so scrolling would not be required to see the full SP table. Additionally, a border was removed to make the transition between the iframe and the SurveyMonkey interface seamless. The HTML iframe coding used was:

\[
<iframe width="890" height="460" src="reference web address" frameborder="0"></iframe>
\]

It is important to note that these were modified for the most recent Chrome, Internet Explorer, and Firefox web browsers on a laptop/desktop computing interface. Other web interfaces, such as mobile web browsers, may not see the intended effect of the iframe attributes.

To match a respondent’s response with the right HTML table, question piping was the main tool used. The underlying logic is to coordinate the response collected with part of the web address that would reference the right HTML code. From the hierarchy of file storage shown in Figure 5-1 above, the destination city’s folder stores multiple folders of locations within the city (in this case the 19 boroughs in Montreal), which stores the specific access and egress information for each origin destination in the Greater Toronto Area. In the proposed survey, the respondent is asked about his or her home FSA address [Q1] the destination of the hypothetical stated preference trip to Montreal [Q2]. When the HTML tables are presented in the stated preference

---

10 This denotes a piping tag similar to the ones used in SurveyMonkey
portion of the survey, the web address of the respondent’s relevant origin/destination HTML table would be:

http://logaday.org/billy/Montreal/[Q2]/[Q1]_1.html

If the respondent had indicated that his or her home FSA was M5T and chosen Montreal destination was Ville-Marie, then the iframe in SurveyMonkey would reference the content found in the web address:

http://logaday.org/billy/Montreal/Ville-Marie/M5T_1.html

In the building phase of the survey, the original naming convention of the html files integrated the destination (Montreal borough) as a number value after the FSA address. This naming convention was initially done to streamline the process the identifying destination locations; however, there was a limitation of the question piping feature of SurveyMonkey where the exact phrase of the response would be piped. If the destination location was indicated as a number value, the corresponding question in SurveyMonkey regarding destination location choice could only be that number value for question piping to work correctly. An alternative tested was to create a separate descriptive text element listing the Montreal boroughs; however, it was problematic to place the text right next to the question. From the limitation, it was found that folders could be created on the logaday.org site using non-roman characters from the French alphabet and the previously indicated storage system was adopted.

Through the use of the HTML iframe and SurveyMonkey’s question piping, it is possible to create only six (6) questions in SurveyMonkey to present the 23,370 possible combinations of origin/destination pairs. The six (6) questions correlate with the six (6) stated preference scenarios that are used in the proposed survey design. This simplification process, using the features provided by SurveyMonkey, greatly reduces the complexity of the survey by creating an artificial dynamic database. Additionally, if changes were required, the workload required is greatly reduced compared to the initial alternative of making 23,370 questions within SurveyMonkey.
5.4 Survey Flow

The collective benefits and limitations of the chosen survey instrument contribute to the overall survey flow of the respondent. For the survey designer, the main objective is to find the optimal point between the data collected from respondents and the ease of a respondent to complete the survey. Typically, a survey collecting a large set of information from a respondent may induce physical and psychological fatigue on the respondent, which may lead to the premature termination of the survey. On the other end, if the survey was made to be easy and quick for the respondent, the resulting data collected may be too simplistic for critical analysis. The proposed survey design and build collects a large amount of data; however, the aim is for the respondent to complete the survey in around 10 minutes, which is considered a short amount of time given the complexity of collected data. This goal was achieved by application of: discrete sections, minimizing pages, using logical skips, and easy to interpret text.

Previously described in section 3, there are four distinct sections to the survey design. Each of these sections relate to a different aspect of an individual’s travel pattern and socioeconomic information. In the survey build in SurveyMonkey, the questions in each section are preceded by a short introductory sentence regarding the questions asked. This introduction is presented to the respondent to get him or her into the frame of mind of the upcoming questions. Another function of having distinct sections is for the survey designer to place only relevant questions within each section. If a potential question deviates from the intended purpose of a certain section, it may not be the optimal location for the question. Additionally, creating discrete sections helped the survey designer to assess the progression of the survey. For example, if revealed local travel questions were asked immediately after revealed intercity travel, the progression of questions may induce some confusion on respondents.

In SurveyMonkey, it is possible to place multiple questions within a single page. Rather than having a respondent click a button to advance from question to question, it is possible to group similar questions together on a single page. The main benefit of this function is reducing the time for a respondent to complete the survey without inducing confusion. In addition, having fewer pages presents the illusion that the survey is shorter to the respondent. For example, the majority of the revealed preference intercity questions is presented in one page and only requires additional pages when logically branching into the mode choice made. The grouping of questions
works within the boundaries of the distinct sections previously explained, where new pages are used for different sections to indicate a new set of questions.

As previously explained, logical skips is a useful feature of web-based survey instruments to omit questions that may be irrelevant to the respondent. Compared to traditional pen-and-paper type of surveys where the respondent has to read the conditions of skipping questions and exactly where to skip to, the autonomous nature of web-based survey instruments reduces this added work. Additionally, this may be a benefit to the researcher as well as it reduces the probability that non-applicable data may be filled out by the respondent. By presenting a dynamic respondent-specific version of the survey design, both the respondent and researcher receive benefit in survey flow.

There may be times when the terminologies used are beyond everyday language. This is applicable within transportation-based research where often used terms within the academic realm are not commonly used by the public; however, the definition of the term is too long-winded to be placed within the survey presentation. One example is the use of the term “egress mode”, which is defined as the mode used to travel from an arrival station to the destination. It may be cumbersome to include the definition within a traditional pen-and-paper survey; it is relatively easy to create a link to open a pop-up browser with the definition of the term. This link can either be built as a definition link somewhere on the page or as a clickable link on the term itself. Within the context of the proposed survey, the ability to quickly provide additional information could be expanded beyond a definition and could be applicable to a map of the trip, additional information about a mode, or further information about a destination.

The use of a web-based survey instrument allows the survey designer to introduce features and attributes within the survey that improves the survey flow for the respondent.
Chapter 6
Sample Selection Methodology

6.1 Target Population

Currently, the target population is any individual who lives within the Greater Toronto Area (GTA). While the project scope is for travel within the Quebec City – Windsor Corridor, the target population was limited to the GTA because of limited time and resources. Even though the target population is any individual in the GTA, a more important subsection of the target population is individuals who have previously made non-commercial intercity trips within the Quebec-Windsor corridor with the trip origin located within the GTA. While individuals who have not made an intercity trip within the QWC are not excluded from the target population, it is expected that individuals who have previously made an intercity trip may find better relation to the metrics stated in the stated preference scenarios.

The non-commercial aspect of the target population is to restrict freight-based trips from the study. Freight-based trips are in a different market segmentation that may be dependent on a different set of decision making attributes than non-commercial travel.

In addition, given the restrictions set by the ethics board on the project, the data can only be collected within the GTA. Future expansion of the survey should include individuals who have made non-commercial intercity trips in any city within the defined major economic corridors within Canada (Quebec-Windsor, Calgary-Edmonton, and Seattle-Vancouver); however, this expansion is dependent on the scope of future work.

6.1.1 Population Count

The main reference statistics to validate the collected responses is population and dwelling counts. While population distribution may not vary greatly compared to dwelling counts, it was assumed that the obtained survey responses are typically collected from one individual within a household. With an emphasis of the survey design to model the effects of local accessibility on intercity mode choice, it is important to include a geographically representative sample of respondents from the target population area.
Currently, the most accurate source of information is the 2011 Census from Statistics Canada, which collected population and dwelling counts for reported FSAs across Canada. All household that provided a postal code with the same FSA were grouped together to calculate FSA totals in the subsequent data tables.

With the available FSA information, the counts can be aggregated up to the city level for easier interpretation and graphical representation. The FSA address from both 2011 Census and survey responses can be aggregated up to the city level by cross-referencing with the list created in section 4.1. Table 6-1 below is a summary of Statistics Canada’s 2011 Census population and dwellings count for the cities located within the target population.

<table>
<thead>
<tr>
<th>City</th>
<th>Population</th>
<th>Dwellings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajax</td>
<td>109290</td>
<td>35478</td>
</tr>
<tr>
<td>Aurora</td>
<td>53203</td>
<td>18092</td>
</tr>
<tr>
<td>Brampton</td>
<td>523881</td>
<td>154657</td>
</tr>
<tr>
<td>Burlington</td>
<td>175733</td>
<td>69795</td>
</tr>
<tr>
<td>Caledon</td>
<td>59635</td>
<td>19699</td>
</tr>
<tr>
<td>Clarington</td>
<td>90003</td>
<td>33123</td>
</tr>
<tr>
<td>East Gwillimbury</td>
<td>10484</td>
<td>3635</td>
</tr>
<tr>
<td>Georgina</td>
<td>47377</td>
<td>19745</td>
</tr>
<tr>
<td>Halton Hills</td>
<td>65931</td>
<td>23016</td>
</tr>
<tr>
<td>King</td>
<td>49053</td>
<td>17285</td>
</tr>
<tr>
<td>Markham</td>
<td>305063</td>
<td>94382</td>
</tr>
<tr>
<td>Milton</td>
<td>78910</td>
<td>26157</td>
</tr>
<tr>
<td>Mississauga</td>
<td>713452</td>
<td>242540</td>
</tr>
<tr>
<td>Newmarket</td>
<td>82861</td>
<td>28994</td>
</tr>
<tr>
<td>Oakville</td>
<td>182528</td>
<td>63897</td>
</tr>
<tr>
<td>Oshawa</td>
<td>149858</td>
<td>61384</td>
</tr>
<tr>
<td>Pickering</td>
<td>87730</td>
<td>29480</td>
</tr>
<tr>
<td>Richmond Hill</td>
<td>185541</td>
<td>60169</td>
</tr>
<tr>
<td>Scugog</td>
<td>23635</td>
<td>9013</td>
</tr>
<tr>
<td>Toronto</td>
<td>2615067</td>
<td>1107853</td>
</tr>
<tr>
<td>Uxbridge</td>
<td>16649</td>
<td>6208</td>
</tr>
<tr>
<td>Vaughan</td>
<td>288344</td>
<td>88295</td>
</tr>
<tr>
<td>Whitby</td>
<td>121006</td>
<td>41575</td>
</tr>
<tr>
<td>Whitchurch-Stouffville</td>
<td>33943</td>
<td>12352</td>
</tr>
</tbody>
</table>
With the low expected number of survey respondents in relation to the Census population and dwelling counts, the relative percentage values should be used for interpretation rather than the count values.

### 6.1.2 Prior Intercity Trip

To assess possible weightings required for data analysis, it is beneficial to recognize the existing number of intercity trips. The source of this information can come from previous data collection programs such as the Canadian Travel Survey or Transportation Tomorrow Survey. The total intercity trips made varied between 400,000 to 550,000 daily trips for both directions. The values for intercity trips were found using the cordon counts by MMM in 2006 and the origin/destination matrix from the 2006 TTS data (Data Management Group, 2012). When comparing the two intercity trip rates, there was a fairly large discrepancy between the two values. The larger value (550,000) was from the cordon count where commercial vehicles were subtracted from the total number of vehicles counted. In comparison, the origin/destination matrix from TTS subtracted all work trips from the total number of trips. The main difference between the two data sets is commercial vehicles versus work based trips. While the cordon count information gave a more suitable value for non-commercial intercity trips, the origin and destination matrix from the 2006 TTS provided location-specific travel demand information.

To interpret a reasonable value of the number of intercity trips made in a day, the total trips using both methods were taken into account. From the 2006 TTS data and the MMM cordon count, a total of 475,000 non-commercial intercity trips, originating from the Greater Toronto Area are estimated to be made daily.

This information could be used to obtain weighting values when modelling intercity revealed preference data. However, these cordon counts only provide the number of vehicles crossing boundaries and may not include other factors such as airplane or rail trips and the origin destination of intercity trips.
6.2 Sample Frame

The objective of the sample frame is to reduce the size of the target population to a more reasonable potential sample size while maintaining the representative demographics of the larger target population. The sample frame can be defined by spatial and temporal boundaries. A number of sample frames have been considered for this project; however, recruiting respondents via online social media may not follow traditional sample framing techniques. For example, the quinquennial Transportation Tomorrow Survey from the Data Management Group at the University of Toronto contacts a representative sample of 5% of all households in the GTHA using a list of households from InfoCanada (Data Management Group, 2010). The following section discusses the proposed sample framing and the challenges associated with attempting to capture survey observations within the target population.

The most optimal sample frame would have the spatial coverage of the boundaries into and out of the GTA exclusively. This spatial would include all trip makers crossing the defined boundary of an intercity trip without having to include intracity trips. Given Sonesson’s research into intercity logit models (Sonesson, 2001), the temporal frame would be fairly long in order to capture recurring trip information from trip makers. Under this sample frame, the survey would be administered via roadside stops along the borders of the GTA. Some travelers would be asked to pull over to complete the voluntary survey. While this sample frame benefits from only collecting data from the necessary demographic of trip makers (intercity travelers), there is an inherent physical danger in roadside data collection especially in a major urban center such as the GTA. A similar type of survey was completed in Medicine Hat on four main roadways (Earth Tech Inc., 2005); however, the relatively lower volume of vehicles in Medicine Hat helped with the overall data collection process.

An alternative sample frame would spatially include the entirety of the GTA in the sample frame. With this approach, a variety of sample recruitment techniques would be used to maximize the number of collected data. One way of looking at this alternative method is that each recruitment technique has a constrained sub-sample frame and the union of all sub-sample frames would become the overall project sample frame. Under this sample frame, it would be unlikely to collect survey records from individuals who are in the process of making an intercity trip. Instead, questions asked in the survey design will determine whether or not an individual has
made an intercity trip in a past period of time. Using that past travel information, the collected information from individuals who match the required criteria will be used for data analysis. The negative aspect of this survey framing approach is a chance of over sampling the target population by collecting information on individuals from a certain socioeconomic level rather than a representative sample of the target population; however, this information can still be beneficial for a sub-objective for the project given enough resources.

Given that one of the objectives is to test the effectiveness of web-based survey methods, the use of multiple sub-sample frames is selected for the project. Unlike previous travel-based surveys, where phone lists were obtained from third party consultants (Data Management Group, 2010), the proposed sample recruitment process is dynamically additive. The specifics of this method of sampling are further explained in section 7.3. To validate that the sample frame matches the target population socioeconomic profiles of the respondents should be compared against existing survey and census data.

The anticipated difficulty with the proposed sample frame is to effectively sample intercity trip makers. The defined timespan regarding the previous intercity trip made by the interviewer is from the present time (currently making an intercity trip) to the maximum length of a year (made an intercity trip a year ago). This temporal range is chosen as specific details regarding an intercity trip made beyond the timespan of a year may not be recollected accurately by a respondent.

6.3 Sample Recruitment

Conventional methods of sample recruitment begin with obtaining a phone list (usually from a third party company) with telephone numbers and home address of households that fall within the defined sample frame. Typically, an advanced letter is sent to the household informing them of the purpose and relevant information of the survey. Afterwards, the household is either contacted by phone or supplied a survey package and asked to complete the survey. Typical response rates vary around 25 to 35% using these conventional methods (Shih & Fan, 2007).

For the proposed survey, an alternative method of collecting survey observations will be used. This alternative method relies on collecting survey responses from a number of different collection sources, including multiple online social media outlets. The sections below detail
possible sources of data as well as anticipated advantages and disadvantages with the source. With this method of sub-sampling, there is a possibility of overlapping sample frames. If a particular socioeconomic or demographic segment is oversampled, then a random sample could be filtered for analysis and used later for model validation.

6.3.1 Online Social Media

The proposed method to obtain respondent utilizes the multitudes of available online social network platforms. To potentially collect respondents with a representative socioeconomic and demographic spread, it is necessary to distribute the proposed survey over different social media sources. Each social media source may contain different number of members, have different online activity patterns, which may or may not affect the number of completed responses.

Sampling multiple social media sources can be analogized to a pyramid, where each increasing level up the pyramid adds additional survey responses to the total amount; however, the number of expected respondents at each increasing level is decreased due to a reduced social network reach.

Under this pyramid analogy, the first step would be to distribute the web-based survey using a social network platform that would produce a large number of respondents knowing that a large percentage of these respondents would fit a narrow socioeconomic demographic. Existing social platforms fitting this would be Facebook and Google. The next step would be to distribute the survey onto other networks that may expand to different socioeconomic and demographic groups. For example, university department listservs may include people outside of the author’s personal Facebook and Google social networks; however, there are fewer people on the listserv and the lack of recognition to the author may lower probability that potential respondents would complete the distributed survey.

One potential problem is the metric to measure the response rate from online social media sources. While the typical method to calculate response rate is the ratio between completed responses to the number of invitations sent, online social networks often do not have set member numbers. In some cases, one individual may be in several overlapping social networks and view an invitation to survey several times. While in other cases, an individual may be in a particular
social network but may never get the survey invitation due to muting notifications or having the invitation notice expire.

The proposed sampling technique has an inherent bias of oversampling respondents with a similar socioeconomic profile and under sampling responses from outside an immediate social network reach. Given enough respondents, it is possible to draw a subset from the oversampled demographics; where multiple draws from the same larger subset can be used to validate the model.

6.3.2 Survey Panel

This method is based on using existing panels of respondents from market research companies to administer the web-based survey design. The individuals in these profiles sign up to the market research service and are provided gift certificates or awards for cooperation. The advantages of using survey panels is an essential guarantee of a certain number of completed surveys as the market research companies charge a fix cost per completed result (SurveyMonkey, 2013). If the survey panel claims to have a representative sample of the target population and has a high response rate, then the resulting data from survey panels should be the most representative of the target population when compared to the other sampling alternatives.

The disadvantage of survey panels is the reliance on a third party entity for data collection. Given certain privacy and proprietary barriers, there might be ambiguity about the source and validity of the collected data. In addition, a set dollar value per collected response might hinder the total number of possible observations if more efficient and/or economical methods of sample recruitment are available.

6.3.3 Phone Lists

Phone lists have been a conventional method of obtaining a sample of households. Phone, household address, postal code information is maintained by third party companies. These databases are available for researchers, marketers, or other individuals to purchase. The benefit of phone lists is the availability of cross tabulation with parameters such as area codes. With definable area codes, clustering techniques can be employed to narrow down the sample size. With phone list information, the corresponding address is typically available. An advance letter can be sent to these addresses informing households about the survey. This advance information
helps to legitimize the survey and increase the response rate. The effectiveness of this approach is seen with the MSU case study where an advanced postcard sent before the survey improved the response rate (Kaplowitz, Hadlock, & Levine, 2006).

Disadvantage of phone list based sample recruitment is the additional cost required to purchase the phone/address list and to send out the advance letter. In addition, in relation to conducting a web-based survey, many of these databases do not include emails. Compared to household telephones and addresses, accurate databases of emails are hard to maintain. Without an email list, the web-based survey would have to be preceded with the initial cost of the phone list as well as resource costs with contacting respondents about the web-based survey.

6.3.4 Street Intercept

Individuals could be intercepted in a shopping mall and asked to take the survey either at the mall location or given information to complete the survey at home. To legitimize the survey purpose, it is beneficial to set up a kiosk within the mall. This allows more information to be displayed to the public as well as to establish formality. A shopping mall was chosen as shopping malls are “public” spaces, easily accessible by most demographics of people, with a relatively high density. Typically, companies set up large flashy displays to catch people’s attention. To further public interest, a variety of incentives are offered given that an individual completes a survey at a computer terminal set up within the area. These incentives range from a small prize advertising the company to a chance to win a large “grand” prize.

Although the surveys are completed via computer terminal and can be considered web-based, the collection of completed surveys is still dependent on direct interaction. In addition, the use of computer terminal as a survey instrument can easily be replaced with pen and paper methods. It would also be difficult to ensure that the collected response matches the required socio-demographic profile and surveys in multiple mall locations are required to get adequate spatial coverage. The inevitable requirement for incentives using this approach would raise costs and cooperation with a corporation might result in tradeoffs with the survey design.

6.3.5 General Advertising

Another method of sample recruitment is general advertising either in print or online forms. This method is the most passive method as there is essentially no direct contact with the potential
interviewer. The benefit of using the advertising approach is savings on manpower as the only requirement needed is to put up the advertisements. For print media, advertisements could be placed in strategic locations such as transit hubs where there is a high density of potential viewers. The size of print ads should be restricted to small posters to reduce cost. Online advertisements have more latitude; however, the reach and impact of the ads are less effective than print medium. A prize incentive could be incorporated with the advertisement and spread over relevant streams of media (blogs, twitter feeds, Facebook) to create awareness. The incentive is to help motivate respondents to complete the survey.

The disadvantage of advertising is the variable effectiveness. At times, small advertising campaigns can attain a high level of social awareness while other times large campaigns flop. The resources spent into advertising should be carefully monitored and issued out in phases to assess the effectiveness of one type of advertising before spending resources.
Chapter 7
Survey Observations

7.1 Pilot Survey

To recruit respondents to complete the survey, an email was sent out to the University of Toronto’s transportation listserv with a direct link to the survey. Only one message was sent via email and using this distribution, 16 completed results were collected over the course of 14 days. For the pilot test, there was no incentive offered. Based on the collected information, the average time to complete the survey is estimated to be around 10 minutes. In addition to the 16 completed surveys, 6 surveys were started but incomplete with an average survey completion rate of 72.7%.

The initial section was the revealed preference section that inquired about the respondent’s most recent intercity travel. Respondents were asked if they had previously made an intercity trip within the past year. The results of the pilot test revealed that 68.8% of respondents had previously made an intercity trip within the defined timeframe.

For revealed intercity travel, the results from the pilot survey indicated that there was a fairly even spread of trip makers across all modes. 36.4% of respondents traveled via automobile, 27.3% traveled by bus, 27.3% traveled by rail, and 9.1% traveled by air. These values were expected as the pool of respondents is primarily students or recent graduates with some individuals from engineering consulting companies.

The collected RP data shows a large variation in travel modes, times, and costs. This result was a little surprising given the use of social media as a method of sample recruitment. It was expected that the use of social media would result in a sample of respondents with similar socio-demographic profiles. For example, if a university student distributed a web-link to the survey using his/her social media network, the anticipated respondents would be expected to be primarily university students with similar travel patterns. While the preliminary results of this pilot survey exhibited some signs of this, the amount of respondents using automobiles and making work-based trips was higher than expected.
Given the low response rate, it was difficult to assess statistically significant results from this portion of the survey. Preliminary interpretation of the stated preference data (from face value) reveals that the two major modal choices from respondents were either automobile or bus. With airplane being the least popular for each situation presented.

Looking at revealed daily travel data, 81.3% of the respondents’ primary travel purpose is to go to school, with 18.8% making work based trips, and 0% primarily making non-work or school based trips. Travel time and costs information was also collected. Most respondents traveled more than 15 minutes door-to-door with costs primarily under $3.00. These daily travel attributes are typical and expected of the demographic of individuals from the email listserv.

This result is a large benefit for the design of the full survey. Assuming that 400 completed surveys from individuals who have made intercity trips are required, only 500 completed surveys should be required from trip makers regardless of making intercity trips given the results of the pilot test. In addition, demographic questions about sex, age, and income were collected. The collected information revealed the following results:

- 10:6 Male to Female ratio
- 50% respondents between 18 and 24 years of age
- 25% respondents reported income under $20,000

The reported socio-demographic profile is consistent with the anticipated recruited sample from the distribution list. Full implementation of the survey would benefit most from having the socioeconomic profile the respondents to represent the profile of the GTA. This profile could be retrieved from current Transportation Tomorrow Survey databases.

### 7.2 Final Survey

Using the comments and suggestions from the pilot survey program, a final survey was created in SurveyMonkey for distribution. The following sections provide detail into each data collection method used. In addition, aggregated data trends are outlined to show the full composition of collected data and show any possible trends of interest. Overall, data collection for the final survey design began on January 23, 2013 and ended on May 15, 2013. The total number of completed surveys was 430 between the data collection times.
7.2.1 Collector Sources

This section details the results of each separate collector source in SurveyMonkey during the data collection program. Collectors are individual internet addresses that can be generated to access one online survey. Rather than having just one database of responses, the main benefit of having multiple collectors is the ability to track the number of individuals that have started/completed the survey without compromising the anonymity of respondents. In SurveyMonkey, individualized settings can also be made for different collectors including; cutoff dates, max response cutoff, password protection, etc. While tracking data sources has previously been done for many non-web-based surveys, the automation of multiple collector sources is an advantage.

Given that all of the survey observations have been collected through the SurveyMonkey web-based survey instrument, it is important to assess the advantages and disadvantages of different web-based outlets. For each collector source, the following information will be outline and discussed (given availability of data):

- Anticipated audience
- Potential audience size
- Total number of surveys started
- Total number of completed surveys
- Generalized sample frame
- Potential for overlapping with other collectors
- Discussion of collector

The goal is to critically assess how social media groups differ from each other as well as to strategize possible improvements in future web-based data collection efforts.

7.2.1.1 Facebook – Profile

Facebook has become one of the largest online social networks at the time of writing with millions of active members. The general idea of Facebook is an outlet for individuals to express essentially anything towards everyone on the site or people within that individual’s social network. In this collector source, a web link and preamble was made available to individuals
within the author’s social network. Only individuals who are on the author’s friends list were able to see this information.

- Collector size: 774
- Surveys started: 36
- Completed surveys: 24

The authors’ friends list contains both close friends and acquaintances to the author. Overall, the general demographic of individuals in this collector source are current university students and some new graduates. While more than half of the individuals are from the GTA, there are a number of individuals who are not within the target population. It is assumed that surveys were not completed by individuals outside of the target population. The socioeconomic profile from this collector is mainly university students or new grads and do not represent a representative GTA demographic. In terms of geographic profile, the abundance of university students are expected to skew household locations around the University of Toronto area; however, there may be some non-Toronto commuters.

It is possible that individuals from this collector source may have some overlap with the other Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This Facebook-based collector source has a very limited sample frame range as it targets individuals within the author’s personal social network. A high number of completed surveys were expected from this collector source because of the personal connection between the author and the potential respondents. Compared to the other collectors, there were a high relative number of completed surveys; however, the overall response rate was still under initial expectations. It is important to note that visibility of the web link on other individual’s Facebook feed are dependent on the overall social activity during the time the link was posted as well as background algorithms to determine relevance of activity. These variables contribute to the visibility of the web link to individuals within the author’s social network.
7.2.1.2 Facebook – Skule Group

Similar to the previous Facebook-based collector, a web link and preamble was made available to individuals within a certain social network subset. Facebook allows for the creation and management of interest groups. It is possible to join a group and not have other group members listed as a Facebook friend. The idea of the Facebook group is to share information to other individuals with common interests without sharing more personal items such as photos or private discussions. In this case, the web link was distributed to the Facebook group Skule™, which is the Facebook group for the Engineering Society at the University of Toronto. In this group, there are both current students as well as alumni.

- Collector size: 2,745
- Surveys started: 47
- Completed surveys: 37

Overall, the general demographic of individuals in this collector source are current university students and some new graduates at the University of Toronto’s Engineering department. The socioeconomic profile from this collector is mainly university students or new grads and do not represent a representative GTA demographic. In terms of geographic profile, the abundance of university students are expected to skew household locations around the University of Toronto area; however, there may be some non-Toronto commuters.

It is possible that individuals from this collector source may have some overlap with the other Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This Facebook-based collector source has a very limited sample frame range as it targets individuals within a specific academic department. A high number of completed surveys was expected from this collector source on the assumption that the author’s name is recognizable by individuals in the Facebook group. Compared to the other collectors, there were a high relative number of completed surveys; however, the overall response rate was still under initial expectations. While the large collector size in the group may signify constant activity and views, it also contributes to a faster rate of additional posts, which pushed the web link below visibility by the majority of individuals. In addition, the majority of posts on this Facebook group were
advertisements to either participate in events or other surveys. The abundance of this type of general advertisement, there is the chance that posts on the group is often ignored.

7.2.1.3 Facebook – MBA Group

Similar to the previous Facebook-based collector, a web link and preamble was made available to individuals within a certain social network subset. Facebook allows for the creation and management of interest groups. It is possible to join a group and not have other group members listed as a Facebook friend. The idea of the Facebook group is to share information to other individuals with common interests without sharing more personal items such as photos or private discussions. In this case, the web link was distributed to the Facebook group mba, which is the Facebook group for previous members of a student group at the University of Toronto’s Engineering department. In this group, there are both current students as well as alumni.

- Collector size: 109
- Surveys started: 10
- Completed surveys: 10

Overall, the general demographic of individuals in this collector source are alumni from the University of Toronto’s Engineering department. Compared to the previous Facebook-based collectors, the members in this group are either new university graduates or individuals who have been in the workforce for less than 10 years. In terms of geographic profile, the small collector size is expected to

It is possible that individuals from this collector source may have some overlap with the other Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

Compared to the other Facebook-based collectors, this collector source has a slightly sample frame range as the majority of the members are spread over a number of graduating years and are currently in the workforce. A fair number of completed surveys was expected from this collector source because of the closer-knit community of the student group. Compared to the other collectors, there were a high number of completed surveys.
7.2.1.4 Reddit – U of T

Reddit is a portal of new and popular information on the internet with the Reddit users dictating the popular content through voting of the good and bad articles. When certain posts receive a certain submission score to signify community approval, then those posts are visible to more individuals on Reddit (Reddit, 2013). The submission score is simply the number of summation of users to who voted for the post to be visible (upvotes) and users who do voted down the post (downvotes). As a user-based portal, anyone can submit a link given that they have a Reddit account. One of the key features of the Reddit community is the hundreds of sub-communities that are focused on a specific topic of interest. For any user, it is possible to subscribe to a number of subreddits that appeal to an individual’s interest (Reddit, 2013). For this specific collector, the web link was posted on Reddit’s U of T subreddit. As the name implies, the focus of this subreddit is topics and discussion about the University of Toronto.

- Collector size: 2,314
- Surveys started: 22
- Completed surveys: 15

As the subreddit is subscribed by current U of T university students, the socioeconomic profile from this collector is expected to be biased towards a certain group of people and not representative of the GTA as a whole. From a geographical standpoint, the relatively low number of individuals on the list and the large university student demographic may not accurately represent GTA household locations; however, it is possible that there are individuals who reside outside of Toronto and within the GTA.

It is possible that individuals from this collector source may have some overlap with the other Reddit collector source. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This Reddit-based collector source has a limited sample frame range as it targets individuals within who are either interested in, attending, or have recently attended the University of Toronto. A low number of completed surveys were expected from this collector source because of the lack of established social presence on Reddit prior to web link posting. Compared to the other collectors, the number of completed surveys was beyond initial expectations. Overall, there
were more individuals in the community that upvoted the link than downvote and even initiated discussions regarding the topic of the survey.

7.2.1.5 Reddit – Toronto

Similar to the previous Reddit-based collector, a web link and preamble was made available to individuals within a certain subreddit community. For this specific collector, the web link was posted on Reddit’s Toronto subreddit. As the name implies, the focus of this subreddit is topics and discussion about the Toronto.

- Collector size: 26,248
- Surveys started: 4
- Completed surveys: 3

This subreddit is subscribed by individuals who are interested in or live in Toronto. While there are no exact statistics of the socioeconomic spread of users, it can be assumed that a fairly wide socioeconomic spectrum is subscribed to this Toronto subreddit. From a geographical standpoint, the individuals within this collector source should have a fairly representative distribution of household locations.

It is possible that individuals from this collector source may have some overlap with the other Reddit collector source. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This Reddit-based collector source has a relatively large sample frame range as it targets individuals within who are either interested in, or currently living in Toronto. While there are subreddits for other municipalities within the GTA, this Toronto subreddit does contain news and articles on topics around the GTA. A low number of completed surveys were expected from this collector source because of the lack of established social presence on Reddit prior to web link posting. Overall, there were more individuals in the community that downvoted the link than upvote. The low popularity was expected but based on the number of registered users, more completed results were expected.
7.2.1.6  Skule Nite Intercept

This collector was collected similar to a street intercept based data collection method. Individuals were asked to complete the survey at the conclusion of a photo shoot for a sketch comedy revue.

- Approximate size: 27
- Surveys started: 8
- Completed surveys: 7

The participants for the photo shoot were current university students and the socioeconomic profile from this collector is expected to be biased towards the student socioeconomic demographic and not representative of the GTA as a whole. From a geographical standpoint, the relatively low number of responses and the large university student demographic may not accurately represent GTA household locations.

It is possible that individuals from this collector source may have some overlap with the Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This collector source has a very limited sample frame range as it targets individuals involved with a specific event within a specific academic department. Initial expectations were survey completions from each individual attending the photo shoot; however, the limitation of one computer and a condensed schedule resulted in a lower number of completed surveys.

7.2.1.7  UTEK

This collector was distributed by the author’s acquaintance to the mailing list for the University of Toronto’s Engineering Kompetition (UTEK). The distribution of the web link to the UTEK mailing list was done as mutual favor.

- Approximate size: unknown
- Surveys started: 11
- Completed surveys: 8

As the UTEK email list contains only current engineering university students, the socioeconomic profile from this collector is expected to be biased towards the student socioeconomic
demographic and not representative of the GTA as a whole. From a geographical standpoint, the relatively low number of responses and the large university student demographic may not accurately represent GTA household locations.

It is possible that individuals from this collector source may have some overlap with the Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This collector source has a very limited sample frame range as it targets individuals’ involved with a specific event within a specific academic department. A low number of completed surveys were expected from this collector source.

7.2.1.8 Skulebook

This collector was distributed by the author’s acquaintance to a potential number of different sources including the acquaintance’s Facebook and email listings. The suggested email listings were the University of Toronto engineering faculty’s yearbook listing as well as the Chinese Engineering Students’ Association also at the University of Toronto. The distribution of the web link to the different distribution sources was done as mutual favor.

- Approximate size: unknown
- Surveys started: 26
- Completed surveys: 17

It is assumed that the individuals from the listings are primarily engineering university students, the socioeconomic profile from this collector is expected to be biased towards the student socioeconomic demographic and not representative of the GTA as a whole. From a geographical standpoint, the relatively low number of responses and the large university student demographic may not accurately represent GTA household locations.

It is possible that individuals from this collector source may have some overlap with the Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.
This collector source has a very limited sample frame range as it targets individuals involved with a specific event within a specific academic department. A low number of completed surveys were expected from this collector source. Despite initial expectations, there were a relatively large number of responses from this collector source. The discrepancy between initial expectations and results may allude to variable survey responsiveness based on the individual distributing the link.

7.2.1.9 EngSoc Digest

The EngSoc Digest is venue to post events and news to the current undergraduate engineering faculty at the University of Toronto. In addition to posting events and articles on the digest webpage, a weekly email is sent to the U of T engineering undergrad students highlighting each article.

- Approximate size: approximately 2000
- Surveys started: 122
- Completed surveys: 81

As the EngSoc Digest email list contains only current engineering university students, the socioeconomic profile from this collector is expected to be biased towards the student socioeconomic demographic and not representative of the GTA as a whole. From a geographical standpoint, the large university student demographic may not accurately represent GTA household locations.

It is possible that individuals from this collector source may have some overlap with the Facebook collector sources. The use of single IP limiters in SurveyMonkey is used to reduce the probability of multiple entries from one respondent.

This collector source has a very limited sample frame range as it targets individuals within a specific academic department. A low number of completed surveys were expected from this collector source on the assumption that the students do not access the EngSoc Digest on a regular basis. While the large collector size in the group may signify constant activity and views, the number of views may be low due to the single-function nature of the EngSoc Digest. In addition, the majority of articles on the EngSoc digest were advertisements to either participate in events
or other surveys and the abundance of this type of general advertisement may lead to many articles being ignored.

Initially there was little activity on this collector despite an ad being posted on the EngSoc Digest website. When the weekly digest email was sent to the students, there was a large increase of surveys started as well as completed surveys recorded from this collector. At the end of the data collection period, this email and online advertising based distribution method yielded the most number of complete survey responses for minimal cost.

7.2.1.10 Transport Email Listserv

The transportation email listserv contains the emails of professors, graduate students, and Engineering Science students in the transportation department at the University of Toronto. In addition, this email listserv also contains some industry contacts with engineers and planners in engineering consulting companies in Toronto.

- Approximate size: approximately 50
- Surveys started: 24
- Completed surveys: 20

As the listserv consists of mostly university students and some university professors, the socioeconomic profile from this collector is expected to be biased towards those two groups and not representative of the GTA as a whole. From a geographical standpoint, the relatively low number of individuals on the list and the large university student demographic may not accurately represent GTA household locations.

It is possible that individuals from this collector source may have some overlap with the Facebook collector sources. However, as the author has fewer than five (5) individuals from the transportation department listed as Facebook contacts, the chance for overlap is minimal.

Given the transportation-oriented academic nature of individuals within this collector source, a relatively high number of responses were expected. Comparing the approximate collector size with the number of completed surveys, this ratio is one of the highest compared to all other collector sources. However, the travel decisions of individuals in the transportation department
may not be representative of the target population and has a chance to potentially skew the model estimation.

7.2.1.11 Metrolinx Employees

In a collaborative effort with Metrolinx, the Research and Business Solutions department has offered to distribute the survey link throughout several departments within Metrolinx.

- Collector size: approximately 30
- Surveys started: 11
- Completed surveys: 9

This collector source includes the office-based workforce currently employed at Metrolinx’s downtown office location. While there are no exact statistics of the socioeconomic spread of users, it can be assumed that a fairly wide socioeconomic spectrum could be represented by the individuals in this collector source. From a geographical standpoint, the individuals within this collector source should have a fairly representative distribution of household locations.

Unlike the previous collector sources, this collector is not inherently linked into online social media. As a result, the probability of overlapping with another collector is expected to be minimal if not zero.

Given the transportation-oriented academic nature of individuals within this collector source, a relatively high number of responses were expected. However, the rate of survey responses has been fairly slow and below expected values. It should also be noted that the travel decisions of individuals employed at Metrolinx has inherent biases and may not be representative of the target population.

7.2.1.12 External Source

A number of survey invitations were sent out to external vendors that had previously worked with Metrolinx on market research projects. As the link was distributed by a third party, specific details of respondents in this collector source are unknown.

- Collector size: unknown
- Surveys started: 13
- Completed surveys: 10

It was assumed that the respondents in this collector source are currently employed in the market research business somewhere in the Greater Toronto Area. Due to the low number of responses, it is difficult to assess the demographic spread of respondents.

Unlike the previous collector sources, this collector is not inherently linked into online social media. As a result, the probability of overlapping with another collector is expected to be minimal if not zero.

A large number of responses were not expected; however, the collection of respondents not currently in a post-secondary institution does benefit the overall data set. As the distribution list is unknown to the author, it is difficult to assess whether or not this collector source is representative of the target population.

7.2.1.13 SurveyMonkey Panel

SurveyMonkey was contacted to set up a panel to collect additional respondents. The panel was requested in the beginning of May when assessing the demographics of the collected data through the previously listed collection sources. Two of the main deficiencies seen in the collected data were the concentration of Toronto-based respondents and the over-representation of post-secondary students.

In the panel request, the main objective listed was to collect responses representative of the Greater Toronto Area. There were no qualifiers applicable to the survey as the inclusion of qualifiers would increase the unit rate for each complete response. SurveyMonkey indicated that their panel was based on the Metro Toronto/Hamilton Designated Market Area (DMA) which is inclusive of the Greater Toronto Area.

Invitations were sent out on May 13\textsuperscript{th} and the panel collector was closed on May 17\textsuperscript{th}. Midway through, the SurveyMonkey project manager was notified that additional responses were required due to some respondents completing the survey in an unrealistically short time. As survey panel respondents typically have faster survey completion times, a cut-off time was set at four (4) minutes.
- Collector size: unknown
- Surveys started: 213
- Completed surveys: 179

Data collected through the SurveyMonkey panel was satisfactory in obtaining a representative sample of the Greater Toronto Area. Unlike the other collectors, the age of respondents was not skewed towards that of a typical post-secondary student. As this collector source contains the largest number of completed surveys, the demographics of these observations are expected to be the most spread out.

One of the questions in the survey asks the respondent regarding the source of the distribution link. While this information is not typically provided by the services managing the survey panel, it appears that the links were typically distributed to three main internet sources; Swagbucks network, IMVU online community, and World Golf Tour game. Looking at these three sources, it is hypothesized that the service managing the survey panel operates on a similar platform of sampling multiple online social networks to obtain a demographically representative profile of respondents.

### 7.2.1.14 Summary

Overall, the data collection procedure was satisfactory in obtaining complete survey responses. While the number of total responses was not as high as originally intended, the resources required to obtain the responses were kept at a minimal. Table 7-1 below is a summary for each collector source.

<table>
<thead>
<tr>
<th>Collector Source</th>
<th>Group Type</th>
<th>Approximate Size</th>
<th>Surveys Started</th>
<th>Completed Surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook – Profile</td>
<td>Close social network</td>
<td>774</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>Facebook – Skule Group</td>
<td></td>
<td>2,745</td>
<td>47</td>
<td>37</td>
</tr>
<tr>
<td>Facebook – MBA Group</td>
<td></td>
<td>109</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Reddit – U of T</td>
<td>Distant social network</td>
<td>2,314</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>Reddit - Toronto</td>
<td></td>
<td>26,248</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Skule Nite Intercept</td>
<td>Personal Favors</td>
<td>27</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>UTEK</td>
<td></td>
<td>unknown</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Skulebook</td>
<td></td>
<td>unknown</td>
<td>26</td>
<td>17</td>
</tr>
<tr>
<td>EngSoc Digest</td>
<td>Email listings</td>
<td>~2,000</td>
<td>122</td>
<td>81</td>
</tr>
<tr>
<td>Transport Email Listserv</td>
<td></td>
<td>~50</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>Source</td>
<td>Type</td>
<td>Metrolinx Employees</td>
<td>External Source</td>
<td>SurveyMonkey Panel</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Metrolinx Employees</td>
<td>unknown</td>
<td>25</td>
<td>19</td>
<td>unknown</td>
</tr>
<tr>
<td>External Source</td>
<td>unknown</td>
<td>13</td>
<td>10</td>
<td>unknown</td>
</tr>
<tr>
<td>SurveyMonkey Panel</td>
<td>Survey panel</td>
<td>213</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td><strong>561</strong></td>
<td><strong>430</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 7-1 – Summary of Data Collection Program

There are several metrics that can be used to gauge the relative success of each collector source. Each metric has its own validity and has other explanatory variables that are hard to explicitly measure, such as an individual’s social reach. Some possible metrics are; ratio of completed surveys to collector size, ratio of completed surveys to surveys started, ratio of surveys started to approximate size, completed surveys accounting for social reach and collector size.

Similar to the initial hypothesis regarding the limited reach of social media, it is evident that the relationship between the approximate size of the collector source and the number of completed surveys is not linear. This observation is in line with initial expectations that listing survey requests in social media networks outside of an individual’s own social network yields a lower number or responses due to the reduced familiarity between members in the social network with the individual posting. Especially with a lack of monetary or prize incentive, there is less willingness for individuals to dedicate time on something with little immediate payout.

Aside from the SurveyMonkey panel, collection of responses did not directly incur any costs. If incentives were offered, additional observations would be expected. Overall, a relatively sufficient number of responses were collected through social media and email without incurring the cost of purchasing external email lists or producing and delivering survey invitations in print form.

7.2.2 SurveyMonkey Panel Data Trends

To assess the distribution of observations obtained from the SurveyMonkey panel, it is important to look at the data trends from respondents. These trends are used to understand both socioeconomic and geographic demographics of respondents. The following sections follow the structure of the four separate sections of the survey design. Given a relatively small number of observations, it is important to ensure that the collected data is relatively representative of the target population. Trends in data can be used to compare with existing census information to validate conformity.
7.2.2.1 Intercity Revealed Preference

The first section collected data regarding respondents’ previous intercity travel patterns within the Quebec City – Windsor Corridor. This data can be used to cross reference with existing surveys on Canadian intercity travel in addition to being used for model estimation purposes.

![Figure 7-1 – Frequency of Provincial Travel (Panel)](image)

When looking at the distribution of provincial travel times in Figure 7-1 above, it is evident that the majority of the latest trips have been made within the past three months. This distribution signifies that intercity trips within the QWC are made quite frequently especially when compared to national and international trips. Comparing the frequency between national and international travel, the collected data suggests that the frequency of national and international travel are similar.
Figure 7-2 – Destination of Last Intercity Trip in the QWC (Panel)

Figure 7-2 above summarizes the destination of the respondent’s last intercity trip in the QWC within the past year. The “No Trip Made” bar signifies respondents who indicated that they have not made a QWC intercity trip within the past year. Looking at the results, Montreal had the highest frequency of reported trips, which matches initial predictions during the survey design stage.

![Bar chart showing destination of last intercity trip in the QWC](chart1.png)

Looking at the revealed trip preference of the respondent’s last intercity trip, it is observed that personal, social, and recreational trips had a higher frequency of response than business trips. From this trend, it could be inferred that the majority of individuals will consider these non-business trips during the stated preference questions.

![Bar chart showing purpose of last intercity trip in the QWC](chart2.png)
Rather than look at the aggregated mode choice of respondents, Figure 7-4 illustrates the various mode choices for each revealed preference intercity destination. For all destinations, the private car mode dominates the other choices. Looking only at the revealed trips to Montreal, private car mode had the highest mode share followed by bus, rail, and airplane respectively. This distribution of mode choice is similar to initial estimation and may be useful for model estimation purposes.

7.2.2.2 Stated Preference

![Figure 7-5 – Purpose of Hypothetical Stated Preference Trip (Panel)](image)

Before the stated preference questions, each respondent was asked to input a potential purpose for the hypothetical intercity trip. While this choice did not change the setup for the SP survey, the collected responses is expected to help improve on future survey iterations or designs. It is observed that the distribution of choice emulate the distribution of intercity travel purpose in the previous revealed preference section. There were more records of personal trips compared to the revealed preference question but overall the distribution is similar.
7.2.2.3 Local Revealed Preference

Figure 7-6 – Purpose of Daily Travel (Panel)

Figure 7-6 above is the distribution of daily travel purpose. The resulting trend is similar to initial predictions where work is the predominant daily trip purpose. There is a slightly higher number of school-based trips than predicted; however, there may be correlation with the source of the survey panel respondents.

7.2.2.4 Socioeconomic

The socioeconomic trends of the SurveyMonkey panel respondents should be relatively representative to the target population.
Looking at the distribution of respondents’ age in Figure 7-7 above the spread is fairly even with a slight skew towards the younger ages. This is expected when using an online survey panel to collect responses. With two of the three major panel sources being online games, the resulting age demographic is to be expected. One thing to note is that seniors would be under-represented in the resulting model estimations.

Using various panel distribution sources aimed at different genders, the distribution of genders is fairly equal at a 1:1 male to female ratio. Similarly, the distribution of marital status mostly polarized between the single or married status with few individuals report being divorced or widowed.
Respondents are typically residing in a household of two, three, or four members. Comparatively, household auto ownership distributions follow a similar trend as household size and seem to suggest that most household tend to own either one or two vehicles.
Looking at the trend for household income, there appears to be two peaks; one at 40-49k and another at over 100k. The presence of these two peaks may be caused by the fairly even split between single and married individuals in the survey panel. The socioeconomic based questions in this survey may not be detailed enough to further assess why this trend occurs.

7.2.3 Non-Panel Data Trends

With the data collected from the multiple collector sources, it is important to look at the data trends from respondents. These trends are used to understand both socioeconomic and geographic demographics of respondents. The following sections follow the structure of the four separate sections of the survey design. Understanding these trends can lead to better understanding of the estimated models, which is especially the case if the collected data does not fully represent the target population.

Due to limited space, the graphs that separate each collector source are located in Appendix N. The following paragraphs still discuss variation of data trends between collector sources in addition to comparisons between the panel and non-panel observations.

7.2.3.1 Intercity Revealed Preference

The first section collected data regarding respondents’ previous intercity travel patterns within the Quebec City – Windsor Corridor. This data can be used to cross reference with existing surveys on Canadian intercity travel in addition to being used for model estimation purposes.
When looking at the distribution of provincial travel frequency in Figure 7-13 above, it is evident that the majority of respondents’ last provincial trips have been made within the past three months. This distribution signifies that intercity trips within the QWC are made quite frequently especially when compared to national and international trips. When looking at the distribution of provincial and national travel times, it is evident that the majority of the latest trips have been most recently made in the past half year. Comparing the frequency between national and international travel, the collected data suggests that individuals travel more frequently out of country than within the country.

Figure 7-14 above summarizes the destination of the respondent’s last intercity trip in the QWC within the past year. The “No Trip Taken” bar signifies respondents who indicated that they have not made a QWC intercity trip within the past year. Looking at the results, Montreal had the...
highest frequency of reported trips, which matched initial predictions during the survey design. It should be noted that the “Other” bar included 14 instances of Waterloo as well as other cities not listed in the original choice list.

Compared to the panel observations, the trend is fairly similar with Montreal being the dominant intercity trip destination. Compared to the survey panel, there is a significant amount of “Other” trips listed by respondents as well as more respondents not making an intercity trip within the past 12 months.

Looking at the revealed trip preference of the respondent’s last intercity trip, it is observed that personal, social, and recreational trips had a higher frequency of response than business trips. This result was mainly the capturing of university students through the survey distribution stages. If a more representative sample of the GTA was surveyed, these trip purpose distributions are expected to change.

Compared with the panel responses, there are more social trips made than personal trips. This flip might have occurred due to the lack additional definition on the difference between personal, social, and recreational trips.
Rather than look at the aggregated mode choice of respondents, Figure 7-16 illustrates the various mode choices for each revealed preference intercity destination. For all destinations except for Montreal, the private car mode dominates the other choices. Looking only at Montreal-based trips, the bus mode had the highest recorded mode share. The large bus share may be related to the large post-secondary demographic in this non-panel data set. This distribution of mode choice is similar to initial estimation and may be useful for model estimation purposes.

### 7.2.3.2 Stated Preference

Before the stated preference questions, each respondent was asked to input a potential purpose for the hypothetical intercity trip. While this choice did not change the setup for the SP survey,
the collected responses is expected to help improve on future survey iterations or designs. Compared to the previous revealed preference question, recreational-based trips have the largest share.

7.2.3.3 Local Revealed Preference

Figure 7-18 – Purpose of Daily Travel (Non-Panel)

Figure 7-18 above is the distribution of daily travel purpose. Considering the large number of respondents collected through social networks of predominantly university students, the relatively large percentage of work-based travel is surprising. One possible explanation for this may be students currently on a work term responding that he or she makes daily work trips; however, this cannot be inferred with the data collected.

Compared to the panel respondents, the daily travel purpose of the non-panel respondents is predominantly school-based. This difference is expected based on the survey collector sources.

7.2.3.4 Socioeconomic

All of the following figures listed in this socioeconomic section are disaggregated to each collector source. With the emphasis on the strengths and weaknesses of using multiple collector sources, the following section will allude to how different collector sources may contribute to the overall data.
Looking at the distribution of respondents’ age in Figure 7-19 above, it is observed that a large majority of respondents are between the 18-24 age brackets. Referring back to section 7.2.1, it was correct to predict that Facebook, U of T subreddit, and university groups yielded a predominantly young respondent demographic. Only the transportation listserv and Facebook mba group collected results from a wider age distribution. From all collector sources, there seems to be a lack of respondents over the age of 41, which may be attributed to exclusively utilizing a web-based survey instrument and collecting responses from individuals within the author’s social network.

Compared to the other metrics explored in this section, the distribution of respondent’s age has the greatest variance between the panel and non-panel respondents. The non-panel collection method utilized the social network of the author while the survey panel management service was able to collect with a wider social network reach. This large variance indicates the weakness with collecting data through the use of social networks.
The distribution of genders seems to greatly vary depending on collector source. This variation may be attributed to the individuals in a certain collector group or may also be attributed to the gender of the individual distributing the survey web link. The overall composition of male to female was roughly 5:4.

While the non-panel respondents had a slightly higher male to female ratio than the panel respondents, the slight difference is not expected to affect model estimation.

Potentially related to the distribution of respondent ages, the respondents were predominantly single. Because the distribution of respondent ages does not correspond to that of the target population, marital status distribution is also not expected to match the target population.
Compared to the relatively equal ratio between single and married in the panel response, there is a large single to married ratio for the non-panel data. If marital status is estimated as a statistically significant attribute with the panel data, then there may be errors incurred when validating with the non-panel data.

While there are subtle differences in the distribution of household sizes between the different collector sources, there are less extreme variations compared to the other socioeconomic statistics. Compared with the panel respondents, there are more single person households with the non-panel, which is expected given the large percentage of single post-secondary respondents in the non-panel data set.
The household auto ownership distributions seem to suggest that most households tend to own zero or one vehicle. As with the other household-based metrics, the distribution between different collector sources is similar. The large percentage of zero car owners is expected given the predominant post-secondary student respondents.

Because of the large number of single university students, the distribution of household income reflects on those two socioeconomic attributes. In addition, there are a large number of individuals who have declined to disclose household income information. Looking at the general trend for all collectors, household income is polarized at either under 20k or over 100k. This two-sided trend is expected based on how the post-secondary respondents interpret a household if he or she is living away from the family home during the school year.
Aside from the extreme income values, the rest of the non-panel household income trends are similar to that of the panel respondents.

7.2.4 Validation of Collected Data

Previously discussed in section 6.1.1, the main metric used to assess collected data is population and dwelling counts. While there are other potential metrics to validate the collected responses, the emphasis on local accessibility is the main motivator to use household population and dwelling counts. One of the main concerns during the data collection process was a possibility where outlying cities in the Greater Toronto Area are underrepresented in comparison with Toronto.

With the relatively small number of collected responses compared to the Census population and dwelling counts, the count of responses, population, and dwellings were calculated as percentages relative to the specified list of cities in the survey design. By changing counts into percentages, it is possible to compare the relative distribution of values across the GTA. Figure 7-25 below illustrates the relative spread of collected responses compared to Statistic Canada’s 2011 Census figures. Data tables related to these distribution values are located in Appendix O.
Looking at the SurveyMonkey panel distribution, the overall trend of respondent distribution does follow the distribution of population and dwellings from the 2011 Census. There are a few areas where the panel distribution deviates such as; the larger share of Toronto-based home origins, lower shares for Brampton and Mississauga, and no shares for some smaller cities. Despite these deviations, the 179 observations from the SurveyMonkey panel are satisfactory in representing the target population.

The other non-SurveyMonkey panel observations were also compared as a group. The majority of the observations were skewed towards a Toronto-based origin with more than 25% difference in share compared to 2011 Census data. This group of observations was in line with shares in Markham and Oakville but was missing observations from many of the smaller cities. Despite the larger deviations to Census data, this data set could be used for further analysis of Toronto-based intercity trips rather than GTA-based.
7.3 Survey Observation Summary

Overall, the responses collected from the final survey were satisfactory given the limited time and resources required to approach different interviewees. From the survey panel and non-survey panel collection approaches, it is evident that the sole use of online social media networks would bias towards certain socioeconomic and geographical demographics. The combined use of both survey and individual collection methods were observed to be adequate to obtain a data set that could be used for demand modelling.
Chapter 8
Data Analysis

8.1 Introduction
With the collected data from the different groups of interviewees, various mathematical approaches can be used to best estimate an empirical mode choice model from both the revealed preference and stated preference responses. The sections below discuss both the modelling frameworks used in deriving an intercity mode choice model as well as the results of the data analysis.

8.2 Modelling Framework
Referencing with previous attempts at choice modelling, it is evident that McFadden’s logit model has been the main framework to understand decision processes. From the literature review in section 2.2, it is evident that the multinomial logit (MNL) and nested logit (NL) models have been used either as the primary demand modelling framework or as baseline models for alternative modelling techniques. The basis of these discrete choice models are based on random utility maximization (RUM) theory. This theory assumes that an individual’s choice, from a number of alternatives, maximizes that individual’s utility (or benefit) within rational behavior.

In RUM based discrete choice theory, the utility of a choice alternative “i” in the choice set “t” \((U_{it})\) is based on the combination of a systematic component \((V_{it})\) and a random component \((\varepsilon_{it})\). The systematic component is the observed deterministic explanatory variables used to measure demand while the random component is an unobserved stochastic error component.

Following the functional form of a linear equation \(y = bx + a\), the utility of either the revealed mode choice in RP surveys or the stated mode choice in SP surveys is a dependent variable \(U_j\) of a set of independent explanatory variables \((\beta x_{ij})\) and an error term \(\varepsilon_j\). This functional form can be written out as:

\[
U_j = (\beta x_{ij})_j + \varepsilon_j
\]
Where \( \beta_j \) is the coefficient for the independent explanatory variable \( j \), \( x_{ji} \) is the \( i \)th observation of the independent explanatory variable \( j \), and \( \varepsilon \) is the error term associated with the defined utility. The proposed functional form is inclusive of linear \( x_{ij} \) variables as well as dummy \( x_{ij} \) variables and log-linear \( x_{ij} \) variables following the function \( \beta_j \ln(x_{ij}) \).

8.2.1 Multinomial Logit

McFadden originally developed the multinomial logit model (MNL) based on Luce’s axiom of independence of irrelevant alternatives. This MNL model assumes an underlying Gumbel distribution and a random sample that is independent and identically distributed (Ashiabor, Baik, & Trani, 2007).

The main difficulty in calibrating the multinomial logit choice model is the associated explanatory variables in the data set. With the defined choice set, the available explanatory variables can be considered generic or alternative specific. Generic variables imply that corresponding beta variables have similar effect on the resulting choice. However, with the multinomial logit model, it is assumed that the effect on one choice alternative would have a positive or negative effect on the other choice set alternatives. Rather than the absolute value of utility, the relative nature of the multinomial logit model should be considered when evaluating the calibrated results.

Given the alternative choice “\( i \)” and the choice set “\( t \)”, the probability of an observed individual choosing alternative “\( i \)” over non-chosen alternatives “\( j \)” follows the equations:

\[ P_{it} = P(U_{it} \geq U_{jt} \text{ for } j \neq i; i, j \in C_t) \]

\[ P_{it} = P(V_{it} + \varepsilon_{it} \geq V_{jt} + \varepsilon_{jt} \text{ for } j \neq i; i, j \in C_t) \]

\[ P_{it} = \int_{-\infty}^{V_{it} - V_{jt}} f(\varepsilon_{jt} - \varepsilon_{it})d(\varepsilon_{jt} - \varepsilon_{it}) \]

A more simplistic approach to assessing the probability of an alternative is to ratio the utility of the alternative “\( i \)” with the sum of all utilities within the choice set “\( t \)”.

\[ P(i|\text{choice set}) = \frac{e^{V_i}}{\sum e^{V_t}} \]
8.2.2 Nested Logit

The nested logit model follows a similar functional form with the multinomial logit model, with the exception that there are multiple levels of conditional choices included in the model in the form of “nests”. As stated, the alternative set lower in the nest are conditional to the higher up alternative set. One of the main reasons to use nested logit models is to overcome the IIA problem of multinomial logit models. Another reason might be to include the multidimensional nature of some choice sets.

As the more disaggregated choice set is conditional to the more aggregated nests, the probability of choosing an alternative “i” will equal to the probability of choosing the alternative “i” conditional to choosing some subset “k”. The full equation to determine the probability of choosing alternative “i” is:

\[
P(i|\text{choice set}) = \left[ \frac{e^{u_iV_i}}{\sum e^{u_jV_j}} \right] \cdot \left[ \frac{\frac{u_i}{\sum e^{u_kV_k}}}{\frac{u_i}{\sum e^{u_kV_k}} + \frac{u_j}{\sum e^{u_pV_p}}} \right]
\]

The above equation was only for a single level nested and additional levels of nesting would include further computational complexities. Compared to the similar equation for multinomial logit model, the functional forms are similar; however, there is the added portion to relate the upper nesting levels.

8.3 Pilot Survey

In the first stage of survey deployment, the University of Toronto Civil Engineering Transportation Demand Management email listserv was used to email potential respondents regarding the proposed survey. From this listserv, there were 22 individual respondents who attempted the survey. From the set of 22 responses, 16 respondents fully completed the survey. For the following modelling and analysis, only the 16 completed responses were used.

The program Biogeme version 2.0 was used to calibrate the proposed multinomial logit model. This program was used for its relative ease in coding and management of data set files. The inputs required within Biogeme are the utilized beta values, the utility formula, defined expressions, and the model used for calibration. As this initial data set was small, low t-Stat values were expected from the Biogeme output; however, the signs of the output coefficients were of importance to validate the survey design and modelling efforts.
The chosen model calibration was determined in conjunction through multiple trials of using various numbers of attributes as explanatory variables. Due to a low number of respondents, the list of attributes has been reduced to overall travel cost and travel time in addition to each alternative’s intercept coefficient. Table 8-1 below shows the results of the regression output:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>t-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car Intercept</td>
<td>0.875</td>
<td>0.310</td>
<td>2.82</td>
<td>0.00</td>
</tr>
<tr>
<td>Bus Intercept</td>
<td>1.03</td>
<td>0.303</td>
<td>3.40</td>
<td>0.00</td>
</tr>
<tr>
<td>Rail Intercept</td>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air Intercept</td>
<td>-1.61</td>
<td>0.632</td>
<td>-2.54</td>
<td>0.01</td>
</tr>
<tr>
<td>HSR Intercept</td>
<td>0.125</td>
<td>0.360</td>
<td>0.35</td>
<td>0.73</td>
</tr>
<tr>
<td>Main Travel Time</td>
<td>-8.73 e-15</td>
<td>1.80 e+308</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Main Travel Cost</td>
<td>-9.66 e-15</td>
<td>1.50</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 8-1 – Results of Biogeme multinomial logit model

The preliminary multinomial logit model makes intuitive sense when compared with initial prediction of results. The intercepts for the travel modes are considered alternative specific while the main travel time and main travel costs are considered generic. The resulting coefficient from the alternative specific mode intercepts are relatively intuitive and in line with initial predictions. As the coefficient values for the generic variables are so small, it is possible to infer the utility of each mode by its intercept. Using the rail mode as a baseline, the modelling efforts revealed that the high speed rail mode has a slightly higher utility. The car and bus modes also have a higher attraction compared to the rail mode. However, intercity travel by air has a much lower utility when compared to the other alternatives. This distribution of utilities may be attributed to the demographic profile of the email list that the pilot test was used to contact potential respondents.

It can be interpreted from the negative main travel time and main travel cost coefficients that an increase in either travel time or travel cost negatively impacts the utility of a respondent choosing each mode. Although the signs for main travel time and main travel costs are correspond with initial predictions, the values of the coefficients are much smaller than anticipated. One possible explanation to the small coefficient values may be attributed to the fact that the varying travel time and travel costs for the different travel modes are already captured in the intercept variable in each mode. In addition, the small number of respondents (19) in the pilot
test may not have induced the variations in the stated preference survey that would reveal the elasticities within each travel mode. However, the ratio between travel time and travel cost can be calculated to obtain a monetary value per additional hour of travel.

\[
\frac{\beta_{Tr} + Scaled_{Tr}}{\beta_{Tr} + Scaled_{Tr}} = \frac{-9.66 \times 100}{-8.73 \times 10} = $11.07/hr
\]

The resulting ratio is comparative with the hourly cost of travel for both auto and common carrier modes if air travel was not included. If more respondents had chosen air travel, this cost per hour would be expected to increase.

### 8.4 Final Survey

The SurveyMonkey panel responses were used for the following model estimations. This data set was used because it was the data set that best matches the demographic profile of the target population. While a data set of 179 observations is not optimal, the use of stated preference data may reduce the number of observations required for statistically significant model estimation.

Additional observations were collected by using multiple distribution sources. From these various collector sources, there were 348 respondents who started the survey. From the total started surveys, 251 respondents fully completed the survey. As this set of 251 responses is not as representative to the target population compared to the survey panel, these observations will be used for model validation purposes. If required, this additional data set could be used in addition to the SurveyMonkey panel observations for model estimation purposes.

As small generic attribute coefficients were obtained in estimating the pilot test data, a different statistical analysis program was tested. The program Stata version 10 was used to calibrate three different demand models. Stata is a relatively intuitive program for statistical analysis. Rather than relying on external data and code files, the interface allows for quick manipulation of model parameters. In addition, a database management system is coded into Stata, allowing for manipulation of the data set using Stata’s programing language. This manipulation was found to be useful to convert existing data sets into a usable format for both multinomial logit and nested logit models. The coding used in Stata for the three model estimations is listed in Appendix P.
The main data manipulation required for both MNL and NL estimation in Stata is the expansion of each stated preference choice scenario to account for the five (5) mode choice alternatives. In the expansion process, the following information is required:

- An identification number that ties each mode choice alternative with the stated preference scenario for each individual
- A name and/or numerical representation of each mode choice alternative
- A choice binary value that indicates the respondent’s chosen alternative

This expansion is necessary when modelling stated preference data. In stated preference, the explanatory variables of the non-chosen modes are also used in the estimation procedure alongside the explanatory variables of the chosen modes.

The chosen model calibrations were determined through multiple trials of using various numbers of attributes as explanatory variables. The sections below describe the modelling procedure for both multinomial logit and nested logit demand models. Each section outlines the model structure used by Stata, additional data set modifiers, explanatory variables used, and potential limitations in the estimated model.

### 8.4.1 Multinomial Logit

To estimate a multinomial logit model using the expanded stated preference data in Stata 10, an alternative-specific conditional logit (asclogit) function was used. This conditional logit model fits McFadden’s choice model, which is a specific case of the more general conditional logistic regression model (ChoiceMetrics, 2012).

The asclogit function allows for both alternative-specific and case-specific variables. Alternative-specific variables vary across both cases and alternatives while case-specific variables vary only across cases (ChoiceMetrics, 2012). The alternative-specific logit model was preferred over the more general conditional logit model. In conditional logit, each row of data is considered an observation with no attribute that groups multiple observations into one case. In the scope of this stated preference survey, each scenario has five observations related to each mode choice alternative.
Various multinomial logit models were estimated in Stata 10. In each trial, different combinations of explanatory variables were tested and checked for resulting significance. Table 8-2 below is the output where intercity demand is estimated to a satisfactory degree and there are an adequate number of explanatory variables in the model estimation.

Overall the estimated multinomial logit model is satisfactory in predicting intercity mode choice. The key explanatory variables (travel time, travel cost, access time, and egress time) were all statistically significant with the expected signs. Additionally, some of the access and egress mode dummy variables were also statistically significant and had intuitive signs.

In this model formulation, only the attributes that were common to all alternatives within a case, such as the socioeconomic attributes, could be modeled as alternative specific explanatory variables. With local access and common carrier based attributes treated as generic variables, the alternative-specific conditional logit formulation in Stata 10 estimated non-zero coefficients for some generic attributes for the automobile mode. Even though an intercity trip by automobile would have no access time, the resulting automobile utility function should not include these non-automobile attributes. Additionally, the inclusion of the null attributes of the automobile attributes may affect the accuracy of the resulting utility of common carrier based attributes. To mitigate against this error, local access explanatory variables should not be considered or another model type should be explored.

Looking at the travel time and travel cost explanatory variables, it is evident that both attributes are negatively affected with increased time and cost respectively. To assess the difference between these two attributes, it is necessary to refer back to the data set used in model estimation. Travel times ranged between 0.82 to 9.69 hours while travel costs ranged between 2.7 to 33.8 ten-dollars (price in dollars divided by 10), resulting in an approximate factor of 3.5 between the two data ranges. Accounting for the both the estimated coefficients and data ranges, respondents placed slightly more utility towards travel cost than travel time.

Between access time and egress time, both access and egress times were estimated as statistically significant. It is important to note that the disutility of egress time was higher than that of access time, which suggests that respondents were more sensitive to long egress trips per unit time. This difference between local access and egress utility was not initially expected.
In addition, there is negative utility placed on accessing a departure station by taking local transit, which is intuitive to predictions considering the additional travel time incurred during transit trips versus other local access modes. For egress modes, a transit trip has a positive utility while a taxi trip has negative utility. The difference in signs between transit access and egress modes should be explored in greater detail if similar trends are observed in the other model estimations.

Some socioeconomic attributes were found as statistically significant attributes to this intercity mode choice model. In Stata 10, choosing automobile as the base alternative yielded the best results. From the multiple trials, it was found that income and age contributed to an individual’s intercity mode choice with younger individuals preferring common carrier modes.

### 8.4.2 Nested Logit – Single Nested

To estimate a multinomial logit model using the expanded stated preference data in Stata 10, a nested logit (nlogit) function was used. This nested logit function performs a full information maximum-likelihood estimation using parameterization consistent with random utility maximization (ChoiceMetrics, 2012). Compared to the previous multinomial logit model, a nested logit model relaxes the assumption of independently distributed errors and the independence of irrelevant alternatives by clustering similar choice alternatives into nests (ChoiceMetrics, 2012).

The chosen nested structure functions to separate the automobile mode from common-carrier modes. By creating these two groupings, the utility of local access/egress attributes for common-carrier modes should not be affected by lack of automobile-based accessibility attributes. Figure 8-1 below is an illustration of the nested logit structure.
Various nested logit models were estimated in Stata 10. In each trial, different combinations of explanatory variables were tested and checked for resulting significance. Table 8-2 below is the output where intercity demand is estimated to a satisfactory degree and there are an adequate number of explanatory variables in the model estimation.

Overall the estimated nested logit model is satisfactory in predicting intercity mode choice. Key explanatory variables (travel time, travel cost, access time, and egress) were generally statistically significant with the expected signs. Like the previous MNL model, some of the access and egress mode dummy variables were also statistically significant and had intuitive signs. A trend was observed where one access mode for the slow nest was statistically significant with low p-values but had high p-values for the fast nest and vice versa.

By creating a nested level to group the possible choice alternatives, it was possible to estimate alternative specific attributes that was not possible in the previous multinomial logit model. For example, variable access time for each choice alternative within a stated preference scenario could not be estimated as an alternative specific attribute using the asclogit function in Stata 10. In comparison, the nesting level generated by the nlogitgen function can be estimated using access time and output alternative specific results based on the nesting groups. Using a nested logit function, the coefficient of access and egress time varied for the different nest groupings (slow and fast common carrier modes) with the additional option to set auto mode as a base alternative. By setting the automobile nesting group as a base outcome, the automobile-based local accessibility coefficients were zero, which is intuitive to the logic that automobile-based intercity travel modes do not require local access and egress. Compared to the multinomial logit model, the ability to estimate variable local accessibility attributes is an improvement in the nested logit model.

Looking at generic travel time and travel cost explanatory variables, it is evident that both attributes are negatively affected with increased time and cost. Compared to the multinomial logit model, the nested logit model estimated a slightly larger emphasis in travel time over travel cost.
Between access time and egress time, both access and egress times were with intuitive signs. The p-value of the slow (bus and rail) modes was higher than expected. The nested logit model estimated different coefficients for the different nest groupings – auto, slow (bus and rail), and fast (air and HSR). Like the previous MNL model, respondents had a higher disutility to egress over access per unit time. In addition, results concluded that respondents were most sensitive to local access and egress times when traveling via air and high speed rail modes. This difference in utility between slow and fast common carrier modes is expected as individuals travelling on the slower bus and rail may have different sensitivities to local travel compared to individuals travelling on the faster air and high speed rail modes.

In terms of access and egress modes, there is a negative utility placed on accessing the bus or rail departure station by transit, which is intuitive to predictions. The utility of accessing air and high speed rail departure stations by transit was also negative but with a smaller coefficient. Similar trends were observed for the transit and taxi egress modes.

Some socioeconomic attributes were found as statistically significant attributes to this intercity mode choice model. In Stata 10, choosing automobile as the base alternative yielded the best results. From the multiple trials, it was found that income and age contributed to an individual’s intercity mode choice.

8.4.3 Nested Logit – Double Nested

The previous nested logit structure separated the automobile mode with common-carrier modes. While this single nested logit structure estimated satisfactory results, further grouping was attempted to create distinction between common-carrier modes. Similar to the single-level logit model, this dual-nested logit function performs a full information maximum-likelihood estimation using parameterization consistent with random utility maximization (ChoiceMetrics, 2012). Compared to the previous multinomial logit model, a nested logit model relaxes the assumption of independently distributed errors and the independence of irrelevant alternatives by clustering similar choice alternatives into nests (ChoiceMetrics, 2012).

For this model, an additional nesting level was created to further group the choice alternatives. Based on the possible explanatory attributes collected from the stated preference survey, a few types of level groupings were explored. The first type was to group common-carrier modes by
the speed of travel; bus and rail are categorized as slow while airplane and high speed rail are categorized as fast. This speed-based grouping was created on the assumption that the utility of local access and egress times are related to the station-to-station travel time.

A second type of grouping was also explored. In this model, common-carrier modes were grouped by the relative location of departure stations. In this case bus was placed in its own group, rail, and high speed rail were categorized as rail while airplane was categorized as air. This model structure was created based on the results from the previous multinomial and nested logit models, which showed access time as a statistically significant predictor. From almost all GTA origin locations, the travel time (for one local access mode) to the bus station at Toronto Coach Terminal may be different to that of the rail and high speed rail station at Union Station and to the airplane departure station at Pearson International Airport. This difference between local access time is the main motivation behind this nesting structure. Figure 8-2 below is a diagram of the double nested structure used in model estimation.

![Double-Nested Logit Structure](image)

Various nested logit models were estimated in Stata 10. In each trial, different combinations of explanatory variables were tested and checked for resulting significance. Table 8-2 below is the output where intercity demand is estimated to a satisfactory degree and there are an adequate number of explanatory variables in the model estimation.

Compared to the previous two model structures, the estimated double-nested logit model assumes that access and egress travel times are perceived similarly for all intercity common carrier modes. These key explanatory variables (travel time, travel cost, access time, and egress
time) were all statistically significant with the expected signs. In addition, some of the access and egress mode dummy variables were statistically significant and had intuitive signs.

Looking at the travel time and travel cost explanatory variables, it is evident that both attributes are negatively affected with increased time and cost. The double nested logit model estimated similar coefficients when compared to the multinomial logit and single nested logit models.

Like travel time, it can be assumed that local access and egress travel times are perceived equally for all travel alternatives. Using the double-nested logit structure, it is possible to estimate generic access and egress time coefficients for the common carrier alternatives while being able to separate coefficients for access and egress modes. Access and egress travel time coefficients were similar with a slightly higher disutility placed on egress time. When compared to intercity travel time, the utility of local access and egress times are between one to two times higher, this is logical as intercity travel times are typically longer by the factor of a number of hours.

In terms of access and egress modes, there is a positive utility placed on accessing a departure station by being dropped off at Union Station and Pearson Airport, which is intuitive to predictions considering the low cost to travel time ratio. Being dropped off at Toronto Coach Terminal was observed to have a negative utility; however, the p-value for this coefficient was quite high. Accessing a departure station by taxi was observed as a positive utility in the mode with accessing Pearson International Airport having the highest coefficient and Toronto Coach Terminal with the lowest coefficient. This difference in coefficient values is intuitive as it roughly matches the ticket price of corresponding intercity travel modes. The significance of two egress modes, transit and taxi, were not initially predicted but is intuitive to the model. As local transit is relatively cheap, it is more likely that trip makers would prefer the transit egress mode over the more expensive taxi alternative.

Some socioeconomic attributes were found as statistically significant attributes to this intercity mode choice model. In Stata 10, choosing automobile as the base alternative yielded the best results. From the multiple trials, it was found that income and age contributed to an individual’s intercity mode choice.
8.4.4 Summary

Overall the double nested logit model used the most logical combination of attributes and grouping to estimate intercity mode choice. The use of two nesting structures separated intercity access and egress modes as alternative-specific while keeping generic access and egress times. In addition, the use of nesting modes separated the auto alternative from the common-carrier modes, minimizing the effect of automobile having no local accessibility measures. Table 8-2 below is a summary of the three different models estimated in Stata 10. The cells shaded indicate p-values that are greater than 0.05, which is an acceptable range given a significance level of 95% used in model estimation. The full output from Stata 10 for the three models is listed in Appendix Q.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Mode</th>
<th>MNL</th>
<th></th>
<th>NL</th>
<th></th>
<th>Double NL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Coeff.</td>
<td>P&gt;</td>
<td>z</td>
<td></td>
<td>Coeff.</td>
<td>P&gt;</td>
</tr>
<tr>
<td>Travel Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>auto</td>
<td>-0.080</td>
<td>0.003</td>
<td>-0.108</td>
<td>0.000</td>
<td>-0.104</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel Cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>auto</td>
<td>-0.082</td>
<td>0.000</td>
<td>-0.044</td>
<td>0.000</td>
<td>-0.046</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>auto</td>
<td>-0.160</td>
<td>0.139</td>
<td>-0.285</td>
<td>0.033</td>
<td>-0.139</td>
<td>0.132</td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egress Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>auto</td>
<td>-0.496</td>
<td>0.003</td>
<td>-0.319</td>
<td>0.165</td>
<td>-0.239</td>
<td>0.026</td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access Mode - Drop Off</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>auto</td>
<td>-0.227</td>
<td>0.062</td>
<td>-0.390</td>
<td>0.022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access Mode - Local Transit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>auto</td>
<td>-0.227</td>
<td>0.062</td>
<td>-0.390</td>
<td>0.022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access Mode - Taxi</td>
<td>auto</td>
<td>fixed</td>
<td>0.100</td>
<td>0.580</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>-</td>
<td>-</td>
<td>0.170</td>
<td>0.136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>-</td>
<td>-</td>
<td>0.199</td>
<td>0.114</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>-</td>
<td>-</td>
<td>0.170</td>
<td>0.136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>-</td>
<td>-</td>
<td>0.170</td>
<td>0.136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egress Mode - Local Transit</td>
<td>auto</td>
<td>fixed</td>
<td>0.402</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>0.003</td>
<td>0.697</td>
<td>-0.003</td>
<td>0.987</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>0.645</td>
<td>0.002</td>
<td>0.069</td>
<td>0.530</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>0.456</td>
<td>0.013</td>
<td>0.069</td>
<td>0.530</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>0.003</td>
<td>0.987</td>
<td>0.069</td>
<td>0.530</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egress Mode - Taxi</td>
<td>auto</td>
<td>fixed</td>
<td>0.170</td>
<td>0.136</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>-0.220</td>
<td>0.082</td>
<td>-0.442</td>
<td>0.002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>-0.121</td>
<td>0.294</td>
<td>0.159</td>
<td>0.373</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>-0.121</td>
<td>0.294</td>
<td>0.159</td>
<td>0.373</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>-0.121</td>
<td>0.294</td>
<td>0.159</td>
<td>0.373</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>auto</td>
<td>fixed</td>
<td>1.237</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>-1.237</td>
<td>0.000</td>
<td>-1.150</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>-0.257</td>
<td>0.223</td>
<td>-0.667</td>
<td>0.063</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>0.177</td>
<td>0.588</td>
<td>-0.900</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>-0.145</td>
<td>0.567</td>
<td>-0.852</td>
<td>0.018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household Income Between $30,000 and $39,999</td>
<td>auto</td>
<td>fixed</td>
<td>0.813</td>
<td>0.094</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>-0.813</td>
<td>0.094</td>
<td>-1.150</td>
<td>0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>-0.482</td>
<td>0.217</td>
<td>-0.667</td>
<td>0.063</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>-0.263</td>
<td>0.524</td>
<td>-0.507</td>
<td>0.160</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>-0.760</td>
<td>0.079</td>
<td>-0.852</td>
<td>0.018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household Income Between $80,000 and $89,999</td>
<td>auto</td>
<td>fixed</td>
<td>0.753</td>
<td>0.040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>-0.923</td>
<td>0.009</td>
<td>-1.122</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>-0.448</td>
<td>0.144</td>
<td>-0.611</td>
<td>0.029</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>-0.875</td>
<td>0.020</td>
<td>-0.900</td>
<td>0.003</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>-0.483</td>
<td>0.110</td>
<td>-0.698</td>
<td>0.010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household Income Over $100,000</td>
<td>auto</td>
<td>fixed</td>
<td>0.230</td>
<td>0.367</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>0.299</td>
<td>0.192</td>
<td>0.028</td>
<td>0.878</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>0.230</td>
<td>0.367</td>
<td>0.073</td>
<td>0.724</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>0.010</td>
<td>0.973</td>
<td>-0.147</td>
<td>0.565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>0.010</td>
<td>0.973</td>
<td>-0.147</td>
<td>0.565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Between 18 to 24</td>
<td>auto</td>
<td>fixed</td>
<td>1.187</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>1.187</td>
<td>0.000</td>
<td>0.319</td>
<td>0.095</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>-0.118</td>
<td>0.670</td>
<td>-0.128</td>
<td>0.579</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>-0.395</td>
<td>0.216</td>
<td>-0.435</td>
<td>0.075</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>0.083</td>
<td>0.750</td>
<td>-0.270</td>
<td>0.205</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age Between 25-30</td>
<td>auto</td>
<td>fixed</td>
<td>1.576</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bus</td>
<td>1.576</td>
<td>0.000</td>
<td>0.776</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rail</td>
<td>0.512</td>
<td>0.083</td>
<td>0.502</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>air</td>
<td>1.016</td>
<td>0.000</td>
<td>0.869</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hsr</td>
<td>1.118</td>
<td>0.000</td>
<td>0.832</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
While there were some instances where the estimated coefficients had a p-value higher than 0.05, these values were still included in the model for comparative measures. For example, the p-value associated with drop off access mode was higher than 0.05 for bus and air but within acceptable range for rail and high speed rail.

### 8.5 Model Validation

As the double nested logit model demonstrated the better measure of intercity mode choice, the following model validation was completed for this model only. Two types of validation were employed to assess the performance of the estimated model. The results of these two validation efforts are summarized in Table 8-3 below. The first level was to compare the observed stated preference choice with an aggregated probability of estimated values. To calculate the error induced through the modelling process, the sum of probabilities for each alternative ($\sum P_{\text{alternative}}$) was compared with the observed data. The error term was determined by finding the percent error induced by model estimation within the same data set. Minimal errors were found, indicating that the double nested logit model accurately predicts the mode choice of the panel data set.

The second validation used is to generate prediction values for the non-panel data set using the model estimated using the panel data set. This validation measure assess whether or not the panel-based model is equally applicable to another similar data set. Once again, the stated and predicted counts for each mode choice alternative are summed up for the non-panel data set. The error term is calculated by taking a ratio between the total stated and predicted counts for each
mode. The errors for predicting auto and high speed rail mode choice were under 3% and the errors for predicting bus, rail, and air were under 10%. While 10% seems like a high value, it should be noted that the demographic of the non-panel group was predominantly post-secondary students and may have slightly varied modal choice to the SurveyMonkey panel demographics.

<table>
<thead>
<tr>
<th>Mode</th>
<th>MOD Stated</th>
<th>MOD Predicted</th>
<th>MOD Error</th>
<th>VAL Stated</th>
<th>VAL Predicted</th>
<th>VAL Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto</td>
<td>388</td>
<td>381</td>
<td>-1.88%</td>
<td>550</td>
<td>549</td>
<td>-0.26%</td>
</tr>
<tr>
<td>Bus</td>
<td>154</td>
<td>158</td>
<td>2.35%</td>
<td>296</td>
<td>325</td>
<td>9.75%</td>
</tr>
<tr>
<td>Rail</td>
<td>167</td>
<td>165</td>
<td>-1.06%</td>
<td>208</td>
<td>202</td>
<td>-3.10%</td>
</tr>
<tr>
<td>Air</td>
<td>154</td>
<td>155</td>
<td>0.39%</td>
<td>169</td>
<td>155</td>
<td>-8.32%</td>
</tr>
<tr>
<td>HSR</td>
<td>211</td>
<td>216</td>
<td>2.30%</td>
<td>283</td>
<td>276</td>
<td>-2.45%</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-3 – Validation 1 and 2

The calibration of the double nested logit model from STATA 10 is observed to be satisfactory in estimating intercity mode choice. While there are some errors when cross-estimating the two data sets, the errors observed are within a satisfactory range given the demographic differences between the panel and non-panel data sets.

8.6 Value of Travel Time Savings

The coefficient estimates from the three different demand models can be used in other forms of econometric analysis. One such form is the calculation of value of time travel savings (VTTS), which is a cost-benefit analysis that assesses the tradeoffs between travel cost and travel time. The concept of VTTS arises from the idea that the necessity of travel derives from demand for activities and the idea that travel time has a negative demand. Depending on the importance of the activity, individuals place a certain importance to reduce the travel time required to reach the activity and may be willing to pay a higher travel cost (Rischatsch, 2009).

Some reasons why individuals may want to reduce his or her travel time are to; use the time saved to yield a monetary benefit, spend the saved time in recreation or other activities, and to reduce any undesired attribute of travel such as discomfort or fatigue (Rischatsch, 2009).

While there are a number of different methods of calculating VTTS, this project will utilize the coefficients estimated in the previous three choice modelling models. VTTS is calculated by
taking the ratio of the coefficient of travel time to coefficient of travel cost. This ratio is multiplied by 10 as the estimated values of travel costs were divided by 10 to be in the same magnitude as the other attributes. The travel time coefficient did not have to be changed as it was already listed in hours.

For access and egress VTTS, the respective local travel times were divided by the intercity travel cost. There are inherent weaknesses to this approach as the intercity travel cost is independent of local travel time; however, the survey design would have been too complex to include the dimension of local travel costs. Alternatively, the ratio does measure the cost that an individual would make to be located closer to the departure station. Table 8-4 below is a summary of VTTS values for the three different models.

<table>
<thead>
<tr>
<th>VTTS Intercity</th>
<th>Mode</th>
<th>MNL</th>
<th>NL</th>
<th>DNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td></td>
<td>$9.73/hr</td>
<td>$24.48/hr</td>
<td>$22.67/hr</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VTTS Access</th>
<th>Mode</th>
<th>MNL</th>
<th>NL</th>
<th>DNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td></td>
<td>$19.51/hr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rail</td>
<td></td>
<td></td>
<td>$28.09/hr</td>
<td></td>
</tr>
<tr>
<td>Air</td>
<td></td>
<td></td>
<td>$64.57/hr</td>
<td>$30.46/hr</td>
</tr>
<tr>
<td>HSR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VTTS Egress</th>
<th>Mode</th>
<th>MNL</th>
<th>NL</th>
<th>DNL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td></td>
<td>$60.55/hr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rail</td>
<td></td>
<td></td>
<td>$72.30/hr</td>
<td></td>
</tr>
<tr>
<td>Air</td>
<td></td>
<td></td>
<td>$154.22/hr</td>
<td>$52.23/hr</td>
</tr>
<tr>
<td>HSR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-4 – Table of Value of Time Travel Savings

In the NL model, there were different access and egress travel time coefficients depending on the speed of the intercity travel mode. These access and egress VTTS values are intuitive as individuals are willing to pay more to access/egress from higher travel cost intercity travel modes. While the local accessibility VTTS values may seem high, local travel times are much shorter than intercity travel times.

Comparatively, the DNL VTTS values are similar to that of the NL VTTS values. There is a slight reduction in intercity VTTS due to a higher cost coefficient. Unlike the NL model, the local access and egress VTTS values are similar and may be more representative given the lower p-values. The VTTS values from the DNL model are reasonable given the relative travel time of the different intercity mode alternatives.
8.7 Sensitivity Analysis

A sensitivity analysis is conducted to assess how the model responds to varying level-of-service attributes. The primary goal is to assess the modal shifts due to changes to all modal alternatives in addition to change for high speed rail. Focus is placed on the hypothetical high speed rail alternative to align with the main objectives of this study; however, further analysis may be done on the collected data. The following sensitivity analysis utilizes all collected respondents from both the panel and non-panel data sources.

Looking at the Figure 8-3 above of varying high speed rail costs, when the cost of high speed rail is lowered by 10%, and then there is an 11% increase in high speed rail ridership. This increase comes from the 1%, 3%, 4%, and 4% reduced ridership of auto, bus, rail, and air modes respectively. The results from this sensitivity analysis are intuitive given that the common carrier modes tended to lose the highest percentage compared to the automobile mode. There is a linear relationship where the inverse mode shift occurs when the cost of high speed rail increases.
Similarly, another analysis was done where the travel costs for all alternatives were changed uniformly. In this scenario, the purpose is to assess modal shifts given a universal change in prices; such as an increase in fuel costs. In this scenario, the model predicts an increase in bus ridership given an increase in travel cost with slight increases in auto and rail modes. At the same time, demand for high speed rail and air travel lowers potentially due to the higher ticket costs. Given the linear relationship of attributes in the double nested logit model, the trend in this sensitivity analysis is expected to be linear and based directly on the differences of established baseline travel costs.

Figure 8-4 – Sensitivity W.R.T. Changing Costs for All Alternatives
Next, the travel time of high speed rail was varied between ±10% of stated value. When the travel time is decreased, the modal share of all other alternatives also decreases. Compared to the sensitivity of varying travel cost, the observed modal shift is only half of each corresponding percent value of change. Looking at the non-high speed rail modes, the auto mode has the lowest sensitivity, followed by the rail, bus, and air alternatives. The change in demand for air travel is most likely due to similarities that the air mode shares with high speed rail in terms of travel time and travel cost.
When travel time of all mode alternatives are varied, the trend in demand is observed as the inverse of the sensitivity w.r.t. changing costs for all alternatives. When stated travel time is reduced, bus has the highest increased demand followed by rail and auto. Alternatively, high speed rail and air modes have negative demand when travel times are reduced. This relationship is most likely a function of the overall travel time for each mode. As the bus mode typically has the longest travel time, a percent change would result in the largest reduction of travel time for bus compared to the other modes.
Chapter 9
Conclusion

The following sections summarize the key aspects of the proposed project. Each section is discussed individually and the outcomes from each section are related back to the initial project objectives. Some of the limitations encountered during the project are also discussed within this section. The steps to potentially address these limitations are discussed in the future works section.

9.1 Survey Design

Overall, the survey design is satisfactory to collect both revealed and stated preference travel information from respondents. The goal of the combined RP-SP survey design was to collect a variety of travel and socioeconomic information from the respondent. Referencing section 3, the survey collected revealed intercity travel data, stated preference travel information, daily local travel data, and socioeconomic information. The demand modelling aspect of this project focused mainly on the stated preference travel information with some association with socioeconomic information. While only a subset of collected data was used in this project’s demand modelling section, the collection of the four sections is justified as there is utility of using the full set of data in future intercity travel demand research.

For the intercity travel sections, the original idea was to produce a set of stated preference scenario related to a respondent’s last intercity trip. Due to a lack of resources, the design was simplified to a hypothetical one-way GTA to Montreal based trip. The collected data confirmed initial hypothesis that the most frequent GTA-based intercity trip destination was Montreal, which means that the hypothetical situation chosen was the most probable one for a random respondent. Despite simplifying for a single destination city, setup of the survey design is designed for future expansion of additional origin and destination pairs.

A number of additional simplifications were also required due to time and resource constraints. Such simplifications included but were not limited to assuming single common-carrier departure stations, single-person intercity trips, lack of stated preference trip purpose, and exclusion of local automobile travel time during an intercity trip. Addressing these simplifications in the SP
design is expected to provide a more realistic hypothetical trip for the respondent; however, the SP design used in this project still provided relatively realistic trip attributes for the respondent despite the simplifications.

9.2 Web-based Build

Using SurveyMonkey as a web-based survey instrument, it was possible to seamlessly include the intercity SP survey while maintaining a consistent survey flow. The implementation of the survey design using SurveyMonkey’s web-survey instrument was satisfactory in distributing surveys. While comparisons between web-based and traditional pen and paper surveys cannot be made directly, there were some key benefits observed when utilizing a web-based survey instrument.

The main observed benefit of using a web-based survey instrument is the inclusion of skip forwarding that automatically progresses the survey flow based on the respondent’s input. Compared to traditional pen and paper or computer-assisted inputs, the use of skip forwards in SurveyMonkey eliminates the need for the respondent or surveyor to find the next question based on the previous question’s eligibility. This method of skip forwarding also presents the respondent with response options only relevant for him or her, which may reduce survey fatigue and increase survey response time. From the collected data, the average time to fully complete the survey was approximately 15 minutes, which is below the initial estimations of approximately 30 minutes.

A second observed benefit in using SurveyMonkey is the ability to set up and keep track of multiple collection sources. The proposed survey was distributed simultaneously to a number of online social networks; however, each social network was provided with a different web address and the observations were appended with collector source identification. With an emphasis on data collection through social media, it was advantageous to assess the socioeconomic profiles inherent in each social media network.

A key component in the web-based survey build was the integration of Excel as a data management tool. Excel was extensively used as a database for the SP survey. The structure of the Excel database functioned as a complier worksheet looking up information from other reference worksheets based on a specific identification value. The benefit observed when using
Excel is the visual confirmation of data. At a glance, it is possible to assess if the compiled tables are outputting the correct values and formulas can be fixed if necessary. This process may not be limited to Excel but may have other applications in the development of other web-based SP surveys including those using PHP coding language.

There were some limitations during the survey build, which are either inherent in most web-based survey instruments or present in SurveyMonkey. Pre-programmed third party web-survey services typically do not offer advanced options for background database calculations. As a result, the SP tables were all generated beforehand and stored on an external server. Pre-generation of the SP tables was a time consuming process that does not facilitate for dynamic changes to level-of-service attributes without regenerating the full set of SP tables. In addition, there were a few times when the external server was experiencing errors or down for maintenance. When a respondent was completing the survey under these times, the SP tables would not be displayed due to server error, causing confusion and prompting the respondent to close the survey.

Another limitation to third party web-survey services is a lack of communication with other web-based applications. It was only possible to indirectly pass information from SurveyMonkey to an external web application by question piping; however, the inverse was not possible. In this scenario, a respondent would select a location on a map application and the address or coordinates would be input into SurveyMonkey. In the proposed design, online maps were only used as reference material for the respondent.

Despite these limitations, building the combined RP-SP survey in SurveyMonkey is relatively straightforward and easy. The integration of the survey design into SurveyMonkey is a novel approach at displaying stated preference information. The combined use of HTML coding and question piping facilitates this integration. The compromises made between a structurally and visually polished data collection instrument and the inclusion of complex survey elements was balanced.

9.3 Data Collection

There were two primary focuses in the data collection stage; the first focus was to assess the effectiveness of collecting responses using web-based methods and the second focus was to
collect a data set that is representative of the target population. Considering the resources available to complete both tasks, the data collection aspect was satisfactory in obtaining usable results for data modelling.

In the first data collection focus, a procedure to sample multiple online social networks was used. The hypothesis was that online social networks have a distinct demographic spread and sampling from multiple social networks would simulate the overall demographics of the target population. Oversampling of specific demographics was expected given the limitation of utilizing the author’s own social network. From SurveyMonkey, multiple collector links was sent out to a number of social networks including; Facebook, Reddit, and school email lists. The use of multiple collector links was observed to be helpful in assessing the socioeconomic and geographic profiles of the different respondent sources.

Geographically, the collected responses from the non-panel data were Toronto-centric with less representation in outlying cities in the GTA. Socioeconomically, the collected responses from the non-panel data was skewed towards the post-secondary student profile with the majority of respondents within the age of 18 to 24. While the use of online social networks of a single individual was not enough to capture a representative sample of the target population, the process of sampling from multiple online social networks was effective in expanding the overall socioeconomic and geographic profiles of respondents.

The main limitation encountered during data collection was the low response rate (ratio between completed surveys and potential audience). The online social networks that have the largest potential audience were typically the ones where the interviewer has the least impact and audience. This inverse relationship was expected; however, the corresponding number of survey completes was lower than initial predictions.

Given the deficiencies observed in collecting survey responses through the use of online social networks, SurveyMonkey was requested to obtain a set of responses using their panel audience. Due to limitations in resources, fewer than 200 panel responses were collected. Compared to the non-panel data, the socioeconomic and geographic profiles of the panel data were much more similar to the target population.
It was observed that the SurveyMonkey panel respondents were sampled in a similar method to the non-panel data. Rather than reaching out to a single list of respondents, multiple online communities were targeted. Each of these online communities had an inherent bias to key socioeconomic features such as gender and age. The greater success of the SurveyMonkey panel compared to the non-panel sources was the use of incentives to collect responses. In the SurveyMonkey panel, completion of the survey would result in some sort of incentive for the respondent, increasing the likelihood that an individual within an online community would spend time to complete the survey.

The main limitation for collecting respondents from the SurveyMonkey panel is cost. Compared to the negligible costs of collecting the non-panel respondents, the cost per completed survey from the SurveyMonkey panel is a barrier to research projects with low budgets. In addition, the cost would increase if additional restrictions (such as restricting number of respondents from each city) were added into the project scope.

Overall, the collection of respondents from both non-panel and SurveyMonkey panel sources is beneficial to the project objectives. Both methods of data collections has inherent limitations; however, the combination of the two may help offset some deficiencies. The relatively large data set can be used for further analysis beyond the scope of this project.

9.4 Demand Modelling

The primary objective of the proposed project is to estimate intercity mode choice within the Quebec City to Windsor Corridor. Using the data collected Stata version 10 software was used to estimate three demand models; a multinomial logit model, a single-nested logit model, and a double-nested logit model. The results from all three models were satisfactory in predicting intercity mode choice. Of the three model estimates, the double-nested logit model was the model structure that best balanced explanatory variables with prediction results.

The double-nested logit model was preferred as the use of two nesting structures allowed for both generic and alternative-specific intercity travel attributes. However, one disadvantage of estimated nested logit models in Stata is the necessity for each alternative to be included in a nesting group for each level. This limitation forces the automobile mode to have common carrier specific attributes and is only zeroed out by stating automobile as a base alternative.
For model estimation, the SurveyMonkey panel data was used because it was more representative of the target population than the non-panel data. In addition to intercity travel time and cost being statistically significant, local access time was determined to statistically influence intercity mode choice. Generally, there was an increased modal share of common carrier modes if local access and egress times were reduced. From various iterations, only some socioeconomic factors were observed as explanatory variables to intercity mode choice. Validation of estimation results on the non-panel data revealed minimal errors in predicting automobile and high speed rail modes and higher errors in predicting bus and air modes.

Analysis was also completed to find basic value of travel time savings and travel time/cost sensitivity. The results obtained from the two methods provided detailed metrics that were in line with initial estimation with dollar values ranging between 20 to 60 dollars an hour. The VTTS amount for local access and egress were higher than that of intercity VTTS mainly because of the different travel times.

The main limitation encountered during model estimation was grouping the alternatives in a logical format. As the automobile intercity travel mode does not have local access and egress attributes, nesting functions had to be applied to separate automobile from the common carrier modes. If the automobile mode was included in a model structure such as multinomial logit or mixed logit models, then there may be a misrepresentation of some level of service attributes due to the improper grouping. Similarly it was observed that the nested logit and the double nested logit had different grouping structures to produce the best analytical models in STATA, this difference should be studied in further detail.

### 9.5 Future Work

The work done in this project can be further applied to multiple project scopes beyond the one used. As previously explained, the procedures of designing and building the survey can be modified for expansion given adequate resources. Within the project scope, the survey can be expanded to include origins and destinations for all possible trips within the Quebec City – Windsor corridor, which would provide further insight regarding mode choice demand within the QWC. By expanding the possible OD matrix in the stated preference survey, an improved relation between RP and SP information can be calculated.
References


City of Toronto. (2013). *TTC Routes and Schedules*. Retrieved 03 20, 2013, from toronto.ca/open: http://www1.toronto.ca/wps/portal/open_data/open_data_item_details?vgnextoid=96f236899e02b210VgnVCM1000003dd60f89RCRD&vgnextchannel=6e886aa8cc819210VgnVCM1000067d60f89RCRD


http://www42.statcan.gc.ca/smr09/smr09_032-eng.htm


Transport Canada. (2003, February). *Via Rail.* Retrieved from Transport Canada:
http://www.tc.gc.ca/eng/mediaroom/backgrounders-b03-f003ed-5399.htm


w3schools. (2013). *CSS Introduction.* Retrieved 03 29, 2013, from w3schools.com:
http://www.w3schools.com/css/css_intro.asp

