UNIVERSITÉ DE MONTRÉAL

BEHAVIORAL APPROACH TO ESTIMATION OF SMART CARD HOLDERS SOCIO-DEMOGRAPHIC CHARACTERISTICS IN A PUBLIC TRANSPORTATION SYSTEM

ANTOINE GRAPPERON
DÉPARTEMENT DES GÉNIES CIVIL, GÉOLOGIQUE ET DES MINES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE CIVIL)
JUIN 2016

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

BEHAVIORAL APPROACH TO ESTIMATION OF SMART CARD HOLDERS SOCIO-DEMOGRAPHIC CHARACTERISTICS IN A PUBLIC TRANSPORTATION SYSTEM

présenté par: GRAPPERON Antoine
en vue de l'obtention du diplôme de: Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de:

Mme MORENCY Catherine, Ph. D., présidente
M. FAROOQ Bilal, Ph. D., membre et directeur de recherche
M. MARTIN Trépanier, Ph. D., membre et codirecteur de recherche
M. FLORIAN Daniel, M. Sc. A., membre

DEDICATION

To my roomates Isabelle and Tiphaine. To my friend Guilhem Poucin who is currently sitting and working on his own thesis one meter away from me. To my family.

In their own way, they all helped me realize how important it is to challenge myself and to leave my comfort zone. ...

ACKNOWLEDGMENTS

I have to thank so many people here! But every contribution was needed to make this project what it is. First, I would like to thank my supervisor Bilal Farooq and my co-supervisor Martin Trépanier.

I also have to thank all students from the Smartcard Lab who participated every week to conf calls to share our work: Oscar Egu (ENTPE, France), Matthieu Galvin and Antoine Giraud (Ecole Polytechnique de Montréal, Canada), Ricardo Cubillos, Jacqueline Arriagada, Catalina Espinoza and Felipe Hernandez (University of Santiago du Chile, Chile).

I would like to thank the engineers who gave insights from the industrial world: Gauthier Cornu (Thalès), Graeme O'Brian (Translink, Vancouver). To this list I add Teresa O'Reilly from Translink, Vancouver.

Thank you to all professors from the transportation field in Ecole Polytechnique de Montréal. Thank to their class and presence they helped me developed my interests for transportation engineering: Robert Chapleau, Catherine Morency, Nicolas Saunier, Hamzeh Alizadeh and Pierre Léo Bourbonnais.

Thank you to Anne Elise Basque, librarian at Polytechnic Montreal for being so patient and answering all my questions about databases.

This work was possible thanks to the much appreciated collaboration from Société des Transports de l'Outaouais. It provided the required databases (travel survey and smart card data).

RÉSUMÉ

Les systèmes de collecte automatisée des titres de transport sont utilisés dans de nombreuses villes, le titre de transport est le plus souvent stocké sur une carte à puce (CAP). Ils génèrent quotidiennement d'importants volumes de données liées à la mobilité des individus. Il devient très intéressant de disposer de méthodes pour pouvoir utiliser ces données car elles présentent le triple avantage d'être exhaustives, longitudinales et directement liées au réseau de transport en commun. En effet, tous les passagers doivent valider leur embarquement (à l'exception des fraudeurs qui s'octroient s'affranchissent de ce devoir), ces données sont recoltées tous les jours de l'année pour l'intégralité du réseau et elles sont liées à un véhicule et une ligne de bus. Il y a de nombreuses applications développées à ce jour: reconstitution des chaînes de déplacements, étude des typologies de déplacements sur le réseaux, étude de la loyauté des usagers, validation des enquêtes de déplacements, identification des maxima de charges sur chaque ligne, étude de l'adéquation de l'offre à la demande, etc.

Pour des raisons de protection de la vie privée, les données de CAP sont pseudonymisées, c'est-à-dire que les informations personnelles incluant les informations socio-démographiques ne sont pas attachées aux enregistrements. Cela explique qu'il y ait peu de travaux de recherche basés sur des donnés de CAP qui impliquent une approche comportementale des usagers des transports en commun. Les travaux qui analysent les comportements des passagers se concentrent sur les habitudes d'utilisation sans jamais pouvoir fournir d'analyse en profondeur de ces habitudes par manque de variables socio-démographiques. De nombreuses attentes liées aux données CAP nécessitent d'avoir accès à ces informations: validation ou complétion des informations provenant d'enquêtes déplacements, analyse avant-après des nouvelles politiques tarifaires ou de tout autre changement sur le réseau, adaptation du service à la clientèle, etc.

La question de l'enrichissement des données de CAP avec des données d'enquêtes de déplacement a déjà été posée sans que des réponses ne soient apportées. Il s'agit d'un problème très délicat notamment car il s'agit de deux jeux de données évoluant dans des dimensions radicalement différentes: les données de CAP ont une grande richesse temporelle mais une grande pauvreté d'information socio-démographique alors que les enquêtes déplacements présentent la caractéristique inverse. L'objet de cette maîtrise est de proposer une méthodologie permettant d'attacher des informations socio-démographiques aux données de carte à puce et de l'appliquer à un cas concret, tout en considérant les questions éthiques liées à la vie privée des usagers.

La méthodologie qui a été développée repose sur trois hypothèses simples. La première est que les individus accédent au transport en commun à pied et par conséquent qu'ils embarquent dans le réseau de transport en commun à un arrêt situé proche de leur lieu de vie. L'arrêt le plus fréquement utilisé au début de la chaîne de déplacement est défini comme l'arrêt le plus proche du lieu de vie. La deuxième hypothèse est que les caractéristiques socio-démographiques des individus ont une influence forte sur leur choix de mobilités. La troisième et dernière hypothèse est que la tarification en vigueur est parfaitement comprise par les usagers qui se comportent alors en agents rationnels et choisissent la solution la moins coûteuse. La méthodologie nécessite deux étapes préliminaires avant une troisième étape d'enrichissement des données de CAP.

La première étape préliminaire consiste à réaliser une synthèse de population avec la propriété d'être la plus précise possiblement spatialement. Ceci est réalisé à l'aide d'un échantillonage par Chaîne de Markov Monte Carlo (dans notre cas: un échantillonneur de Gibbs) qui permet de fusionner plusieurs sources d'information disponibles à plusieurs niveaux d'aggrégation différents. Les résultats de cette étape sont contrôlés spatialement à l'aide de statistiques traditionnellement utilisées: l'erreur totale absolue (TAE), l'erreur standardisée absolue (SAE) et la racine carrée de la moyenne des erreurs au carré (SRMSE). Cette étape donne des résultats précis spatialement.

La deuxième étape préliminaire consiste à développer et calibrer un modèle de choix discret de chaîne de déplacement à l'aide de l'enquête de déplacement. Dans notre cas, le modèle développé est basé sur une structure emboîtée-jointe. La première couche est le choix du mode et la seconde couche du nid est le choix de la chaîne de déplacement. Le modèle est calibré sur 80% de l'enquête déplacement à l'aide du logiciel Biogeme et il est validé à l'aide des statistiques habituelles et par simulation sur 100% des données de l'enquête déplacement. Ce modèle a de bonnes statistiques et donne des résultats pouvant être fortement améliorés.

L'étape d'enrichissement des données de CAP utilise toutes les hypothèses: les données de CAP sont analysées et préparées afin d'identifier la zone géographique où habite, pour chaque carte, le détenteur de la carte; et afin d'identifier le type de chaîne de déplacement qui a été réalisée. Les chaînes de déplacement sont décrites à l'aide de quatre attributs: la fidélité aux services de la STO, l'heure du premier départ (avant, pendant ou après l'heure de pointe du matin), l'heure du dernier départ (avant, pendant ou après l'heure de pointe du soir) et le nombre moyen d'activités journalières. Ensuite, pour chaque zone géographique, le modèle de choix de chaîne de déplacement est appliqué à la population locale avec pour alternatives les CAP qui ont été localisées dans la même zone spatiale. La solution la plus probable est alors calculée à l'aide de l'algorithme Hongrois.

L'application de cette méthodologie repose lourdement sur le concept de modélisation orientée-objet implémentée en JAVA. Le code est rendu public et disponible en ligne. Ne disposant pas de la possibilité de valider notre méthode à l'aide de la vérité, nous procédons à des validations partielles sur la capacité de la méthodologie à reproduire la population empruntant les transports en commun au niveau de la ville et à un niveau local. Le processus de distribution des cartes à puce à la population est réalisé de quatre façons différentes. Premièrement de façon totalement aléatoire. Deuxièmement en utilisant l'hypothèse de géolocalisation des cartes à puces. Troisièmement en utilisant la localisation des cartes à puce et le modèle de choix de chaine de déplacement. Et quatrièmement en utilisant, en plus, la tarification comme moyen de s'assurer de distribuer les cartes à puce à quelqu'un qui a la bonne catégorie d'occupation. Les résultats montrent une nette amélioration par rapport à une affectation aléatoire d'attributs socio-démographiques et la part de chaque hypothèse dans cette amélioration est estimable. Le modèle de choix de chaîne de déplacement est un élément clef de notre méthodologie, cependant pour notre cas d'étude nous bénéficions d'une enquête déplacements présentant peu d'atouts de compatibilité: une part modale très faible pour le transport en commun et aucune dimension de longitudinalité.

Pour validation, nous vérifions que les distributions des divers attributs (âge, genre, occupation, nombre de personnes dans le ménage, nombre de véhicules) correspondent aux distributions observées lors de l'enquête origine-destination. Au niveau global, la distribution de l'âge est le meilleur exemple de la complémentarité des trois hypothèses appliquées. Par exemple, pour la catégorie d'âge 11 à 19 ans, l'enquête OD prévoit 6 000 passagers de la STO. Une affectation aléatoire des cartes à puce aboutit à 3 200, l'hypothèse de localisation des cartes à puce aboutit à 3 500, le modèle de choix de chaîne de déplacement permet de passer à 5000. Utiliser les trois hypothèses permet de monter à 5 600 passagers. On passe de 50% à 93% de la population réelle. Ceci est l'exemple le plus parlant mais la même dynamique se retrouve pour les autres catégories. Il est à noter que la dynamique est plus forte pour la distribution de l'âge car l'effort de modélisation a principalement porté sur l'âge. Il est impossible de valider notre travail au niveau local car l'enquête déplacements ne fournit pas d'information suffisamment précise. La cohérence de la distribution jointe des attributs socio-démographiques est étudiée à l'aide de la valeur de la racine carrée de la moyenne des écarts aux carrés (SRMSE). On note que seule l'application des trois hypothèses permet de diminuer cette erreur de 2.410 à 2.280.

La méthodologie proposée est complexe et comporte de nombreuses sources d'erreur qui viennent s'accumuler. Chaque élément de la méthodologie peut être contrôlé et l'élément clef de la méthode repose sur une bonne qualité des données, notamment de l'enquête déplacements, ainsi que sur un bon taux de pénétration de la CAP comme titre de transport. La

méthode pourrait être adaptée pour des systèmes de transport où le taux de pénétration des cartes à puce n'est pas optimal au prix d'un modèle de possession de la CAP. En l'état, la méthode a prouvé fournir des résultats encourageants mais nécessite cependant plus de travail au niveau du modèle de choix de chaîne de déplacement et surtout au niveau de la validation. Les sources d'erreur sont principalement a) dues aux données, b) dues aux divers modules utilisés qui apportent chacun son lot d'erreur et c) dues aux hypothèses faites qui sont simplificatrices.

ABSTRACT

Automated Fare Collection (AFC) systems such as smart cards are being used in many different cities and countries. The AFC systems leverage large volume of data related to person's mobility and it becomes very interesting to develop methods to use these data to complement other data sources. They present four main advantages, they are longitudinal, they concern every public transit user (within the limitation of the penetration rate of the smart card and the policy around shared smart card ownership), they are passively leveraged and they are directly related to the public transit structure. There are already many applications such as processing the trip chain, study of public transit users loyalty and behaviour, validation of travel surveys etc.

For privacy concerns, smart card data is anonymised and any socio-demographic data is lost. This explains why there are so few research based on smart card data, which involve a behavioral analysis of public transit users. They usually focus on travel habits and have no explanatory power. There are many expectations for smart card data which require to have access to socio-demographic information.

Enriching smart card data to the smart card owner level has already been highlighted as a key aspect of smart card data theory. However it is a very delicate problem since smart card data and travel survey evolve in two different spaces. Smart card data have very rich temporal information but poor socio-demographic information, while travel survey is the opposite. The goal of the research is to develop a methodology to fuse these two datasets while minimizing the introduction of any bias.

The work is developed based on three hypothesis. First, people access public transit by walking and therefore live in the neighborhood of their main station. Second, socio-demographic attributes are key to mobility choices of individuals. Third, everyone understands the fare policy and chose the most cost efficient solution. The methodology requires two steps of preparation.

First of all we realize a population synthesis using a Gibbs sampler to draw agents directly from the underlying distribution. It allows to fuse several data sets. This step is validated using classical population synthesis validation techniques such as Total Absolute Errors, Standardized Absolute Errors and Square Root Mean Squared Errors. We validate our methodology at the most spatially accurate level: dissemination area.

Secondly we design and calibrate a discrete nested-joint model for trip chain choice. The

model is calibrated on 80% of the data and simulated on 100%. We use traditional discrete choice modeling stats to help use develop our model. The model was estimated using the Biogeme software. Our model has a nested-joint structure. The upper nest is the mode choice and the lower nest is the trip chain choice.

Finally we process smart cards. They are enriched to the level of trip chain by identifying alighting stops and computing daily average statistics such as first departure hour, last departure hour, usage of public transit, number of activities over week-days. Then we assign a synthetic agent to each smart card using the model to find the most probable solution.

The implementation of this methodology relies heavily on the Oriented Object Modeling concept. A Java code was developed and is available online. The methodology was not validated against ground truth (microscopic validation) but against reported truth at the city level (macro validation) and at a more local level (mesoscopic validation). Results show that our methodology provides better results than a random distribution of attributes. The trip chain choice model is a key element to our methodology, but in our case study it was also the main source of errors. Results are validated against distributions. For instance, at the macroscopic level, for age category 11 to 19 years old, the OD survey registered 6 000 STO riders. While giving randomly smart cards to the population, we have 3 200 STO riders (approximately 50% of the real population). While using the spatial location of smart cards, we have 3 500 STO riders. While using, in addition, the trip chain choice model, we have 5 000 STO riders. And eventually, while using the hypotheses, we find a population of 5 600 STO riders (approximately 93% of the real population). This is the most encouraging result, however, similar trends can be found on other categories and attributes. It has to be noted that age category was also a key aspect of the trip chain choice model, therefore we were expecting to get better results for this attribute. The mesoscopic validation could not be achieved because the smart card data and the travel survey data are not mutually consistent at the local scale.

The proposed methodological framework is complex and induces many sources of errors. It could be improved especially by modeling smart card owners from non smart card owners.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ACRONYMS AND ABREVIATIONS

AFC	Automated Fare Collection
AVL	Automated Vehicle Location
CAP	Carte à puce
CMA	Census Metropolitan Area
CPM	Computer based model
DA	Dissemination Area
ENTPE	École nationale des travaux publics de l'état
FAMOS	Florida Activity MObility Simulator
GPS	Global Positioning System
Id	Identifier
ILUMASS	Integrated Land Use Modeling And transportation System Simulation
ILUTE	Integrated Land Use Transport Environment
IPF	Iterative Proportional Fitting
LUTI	Land Use Transport Integrated
MCMC	Monte Carlo Markov Chain
OD	Origin-destination
PUMS	Public Use Microdata Sample
RAM	Random Access Memory
SAE	Standardized Absolute Errors
SRMSE	Square Root Mean Squared Errors
STO	Societétée des transports de l'Outaouais
TAE	Total Absolute Errors
TASHA	Travel Activity Scheduler for Household Agents
TRESIS	TRansportation and Environmental Strategy Impact Simulator

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Automated fare collection (AFC) systems in public transit networks are leveraging important amount of data for ticketing purposes. However these kind of data are currently being used to answer many other problems related to transportation management and planning. However AFC systems were designed for ticketing purposes and the data it is producing lacks socio-demographic characteristics and has to be processed. Actually, one of the key topic related to AFC data is to develop methods to enrich it and make it understandable.

1.1 Motivation of Research

The classic way to accomplish strategic and operational planning for a public transit operator and urban planners is through travel surveys and on-board surveys. They are interesting datasets since they include comprehensive socio-demographic information about the respondents. They are designed to fulfill most of the data needs that operators and planners can have, they have information about all modes, the trip chain constitution and intra-household interactions. For the same area, they are often conducted every five to ten years using each time a similar or compatible methodology, which insure a reasonable consistency of data through time. They can be used to do aggregated or disaggregated analysis, forecasting analysis, policy analysis etc. They are also used to calibrate models. However this kind of survey presents many downfalls. The sampling strategy is traditionally a random sample amongst the population which own a land line. However some part of the population is not reachable through land line, especially young adults who are more likely to own a cell-phone than to have a land line. Also there is a growing increase of non-response rate to surveys due to commercial surveys that saturated the market. In addition to this sampling problem, land-line based survey are getting information from a proxy respondent; usually an adult from the household (mother or father) answering for the whole family which induces an important bias. See Stopher and Greaves (2007),Fink (2012),de Dios Ortúzar and Willumsen (2011) and Bayart and Bonnel (2015) for a more detailed literature about survey's data limitations. Surveys are really expansive and laborious to produce and validate (Chapleau et al., 2008) and they represent only an average day for a sample of the population. Innovative surveying techniques are being developed in order to overcome those issues. For example by using a mix of land-line survey with web surveys (de Dios Ortúzar and Willumsen, 2011) (Bayart and Bonnel, 2010), or by assisting the reporting process with a GPS and a computer (Stopher and Greaves, 2007). However some researchers are wondering if it is reasonable to hope to

fulfill our data needs through surveys (Axhausen, 1998). It seems especially improbable to capture the longitudinality of mobility behaviors since the Mobi*Drive* project of leveraging mobility data over six weeks (see Axhausen et al. (2002)) shown that there is a high rate of people dropping off of the study since it was to exhausting. And this cannot be conducted at city scale.

Automated fare collection (AFC) systems in public transit such as smart cards are now being used in many different cities and countries. The AFC system gathers a massive volume of transaction information and recently, making sense out of this data has become a research hot topic. It is an interesting dataset for three main reasons a) it is not a sample since all the public transit riders are recorded (with respect to the smart card penetration rate and the amount of free riders)(Bagchi and White, 2005), b) it is longitudinal data: it captures the temporal evolution of riders behaviors, and c) it is tied to detailed transit operational data (Bagchi and White, 2004; Chapleau et al., 2008). These three advantages allow researchers to perform various types of studies, among them: finding the maximum load, analyzing public transit users loyalty and behavior, looking for hardware failure through the data being produced, assessing whether the supply is matching the demand in the best way possible, analyzing transfer time etc. However AFC data presents drawbacks as well. A major one is that not every public transit user owns a smart card, some of them may still use tickets paid with money, therefore it is hard and quite impossible to consider them in the analysis, which mitigates the value of the analysis. The penetration rate is an important value to control, while doing analysis using smart card data. To achieve a good potential, the smart card data has to meet some requirements. For each transaction it has to be linked to a fare type, a location, a time stamp and operational information (vehicle ID, route ID). Often transportation system are tap-in only and do not require a tap-out, therefore only trip origins are known. Finding the alighting location was a topic for many research papers (Trépanier et al., 2007; Chu, 2010; Munizaga and Palma, 2012; Nassir et al., 2015) and although the theory is quite strong on this field, it still produces some errors. The other lacking dimension to smart card data is the socio-demographic dimension. For ethical reasons, smart card data is anonymized and it is not likely to change (Cottrill, 2009). Therefore behavioral analysis of smart card data remain to the level of describing the use of the smart card (loyalty analysis, mobility patterns), but do not provide any explanatory power.

To make the best use of AFC datasets, there is a need to enrich them with socio-demographic information (Pelletier et al., 2011) since these kind of attributes are key to understand individual's mobility (Bayart et al., 2009). In some AFC system, the fare policy can be very detailed and offer a good granularity to attach some information (usually: adult, student, senior). In a few cities, personal information about smart card owner was known. For in-

stance Viggiano et al. (2016) had access to home address of smart card owners and they used it to study access time to public transport. Munizaga et al. (2014) had access to a small sample of smart card users socio-economic characteristics and they validated destination and purpose inference methods. To the best of our knowledge, the only work that was done to attach socio-economic attributes to smart cards can be found in Trépanier et al. (2012): for each transit rider, they estimated living neighborhood and attached attribute's spatial average to the smart card holder. There is the underlying assumption that everyone in the neighborhood has the same probability to use public transit. This is not satisfactory since socio-demographic attributes have a significant impact on mode choice. We propose a methodology to infer socio-demographic attributes to smart cards mainly based on the analysis of the smart card data. We choose this approach confidently since destination inference methodologies are rather reliable (Munizaga et al., 2014) and smart card allows to make a rather detailed analysis of public transit riders (Ortega-Tong, 2013). The methodology consists in calibrating a trip chain choice and applying it on a synthetic population with a choice set constituted of trip chains observed through the AFC system.

1.2 Dimensions of the problem

People's mobility can be described at various level of details. The shortest element is the trip stage (Bagchi and White, 2005; Chakirov and Erath, 2012): it is accomplished in a single vehicle (walking being considered as a vehicle), from a boarding location to a destination location. The trip is the succession of trip stages from an activity location to another activity location (Meyer and Miller, 2001; Bonnel, 2002). We consider that waiting for a connection at a bus station is not an activity. The trip chain is the collection of trips accomplished during a day (Meyer and Miller, 2001; Bonnel, 2002). It can be as simple as going from home to work, then going back home; or it can be a complex trip chain including multiple activities and sub-tours. The mobility pattern of an individual is all his/her trip-chain put together over periods of time longer than one day (Adler and Ben-Akiva, 1979). The temporal frame of a mobility pattern has various scales, it can be daily, weekly or seasonal. The interest for weekly mobility patterns rely on the assumption that daily travel behavior presents variation along the week (Cui et al., 2014) because activities can be scheduled on various days of the week. However we expect people to follow a weekly schedule. The interest on seasonal mobility patterns relies on the assumption that travel behaviors changes according to the seasons (Morency et al., 2007).

Our case study is the STO public transit network in Gatineau (Canada), October 2005. This case study is potentially a really good study since the smart card system was already installed

for a few years and it is providing the highest level of information: fare, boarding station, vehicle ID and time stamp. The local travel survey and the Canadian national census were held during the same time window, therefore we can consistently use all this information.

Conceptually, there are four dimensions in our datasets: the time dimension, the spatial dimension, the socio-demographic dimension and the mobility choice dimension. We need to get the best information on each dimension. Information about the Gatineau population is available at various spatial scales and we need to be able to consider the most detailed spatial information while considering less detailed information as well. The enrichment over spatial dimension is reached using the most spatially accurate information from the Canadian census. Census includes a rounding process which may reduce the precision of our process. The enrichment over time is achieved through the fusion with smart card data. The enrichment on socio-demographic dimension is achieved through fusion of the travel survey and census data.

There are already techniques to fuse information from travel survey data and census data, but to the best of our knowledge there is no data fusion techniques able to fuse information from travel survey data, smart card data and census data. Data fusion challenges between transportation related datasets were described in Bayart et al. (2009) and in Venigalla (2004). Examples of data fusion between actively collected datasets were given: fusion of travel surveys conducted with various methodologies (web based surveys, land line based surveys) and fusion of travel surveys with census data. However, to the best of our knowledge, nothing was done to fuse actively collected data (surveys) with passively collected data (smart card data or other). It requires to define a framework and describe the methodology. Our methodology is using many components from well known theoretical area such as population synthesis, activity choice model, smart card data processing and graph theory.

1.3 Research objectives

The goal of this work is about developing further the existing theory on enrichment of smart card and proposing a methodology to attach socio-demographic attributes to smart cards while respecting the due privacy of smart card owners. The global project could be seen as an information fusion approach of multiple datasets evolving in various dimensions.

The first step is to understand those datasets: their structure, their strength, their weaknesses and the sources of errors that are induced. The literature review has to cover those points. A state of the art of the research done using AFC data is important to understand what are classic uses of smart card data and how it can help reaching our goal. Then we have to build

a broader understanding of the mobility data through behavioral description.

The main goal of this research is to develop a methodology to attach socio-demographic attributes to smart cards. This has to be done in the global transportation framework that we have described by assembling various pieces of theoretical knowledge. Therefore it is not expected to use classical information fusion techniques but to come up with one specific to the transportation field. The methodology has to be described using theoretical concepts so it can be applied to any case study. It has to be applied to the specific case study of Gatineau (Canada) and the way to apply it has to be described and understandable. Its benefits/limits have to be expressed and analyzed based on the case studyd.

1.4 Structure of the thesis

The thesis layout is as follow. The state of the art section has many subsections. The first three are describing the AFC data theoretical environment: how the data is leveraged, the latest developments in use of AFC data and the expectations for AFC data. Then the methodology section introduces our methodology: first the context is described, then the theoretical description of the methodology is given, then every methodological model required to reach a practical implementation of the methodology is described: a) enriching smart card data to the trip chain level b) generating a synthetic population c) calibrating a nested-joint model for trip chain choice d) assigning observed trip chains to the synthetic population. The third section presents the results of our experiments which are based on one month of AFC data from the STO. The fourth and last section summarizes and concludes this thesis and discusses the future directions.

CHAPTER 2 STATE OF THE ART

This section covers the literature related to our problem. The first four subsections introduce AFC systems as well as a literature review of uses of smart card data. The fifth subsection gives an overview of activity choice model's variety with a special focus on joint models. The sixth subsection is dedicated to population synthesis.

2.1 Automated Fare Collected data in the public transport system

The use of a smart card or any other medium (smart phone for example) in an AFC system is quite simple: when the rider hops in a vehicle of the system, he validates his own boarding by placing his smart card on a smart card reader. The card checks if the rider is allowed to board, with respect to the faring policy, and register the transaction in the system (Pelletier et al., 2011). There is some heterogeneity among AFC systems and the accuracy of the data that is being produced (Erhardt, 2016). Bagchi and White (2004) identify five dimensions to this heterogeneity: a) time information, b) purchase information (type of ticket, price etc), c) structural information (transportation mode, line, direction), which are systematically being recorded, d) the spatial information which is not systematically being record since it requires that the vehicle fleet is equipped with an Automated Vehicle Location (AVL) system, and e) the rider's personal information which may exist but are never displayed due to privacy concerns.

AFC systems produce data about complete disaggregated individuals mobility. There are some limitations to the accuracy of this kind of data. It can be affected by hardware failures, bus overload situations, driver's mistake etc, and it needs methods to be able to detect them and handle them (Chu, 2010). Many public transit faring systems that include a smart card system also include paper tickets and cash as a way of boarding. This kind of boarding is less interesting since they cannot be linked to one another and the behavioral dimension of trip chaining is lost. Therefore the penetration rate of the smart card in the public transit rider population may be a limitation to the use of the data (Erhardt, 2016).

2.2 Destination and activity inference methods

In some faring system, only a tap-in when boarding is required: the AFC system grants the authorization to board or otherwise. Inferring the alighting has been a major issue for the past few years. The main methodology to do so is a rule-based approach fully detailed in

Trépanier et al. (2007) (see figure 2.1, for further details) which adopts an comprehensive approach of the trip chain structure. The vanishing route (Trépanier et al., 2007) is the sequence of stops where the user can alight after his boarding. For an individual, if the next boarding is close to the vanishing road of the bus he is in, then it is considered that he hops off the bus at the closest station of its next boarding, provided there is one within a walkable range. It has been applied in Gatineau, Canada (Trépanier et al., 2007), Santiago, Chile (Munizaga and Palma, 2012) (with improvements to be able to consider that a bus route can be very tortuous or buckles on itself), Brisbane, Australia (Alsger et al., 2015). This methodology lets some unlinked trips and He and Trépanier (2015) proposed an improvement using a kernel density estimation on the historical data of the smart card. This idea of using travel habits to infer more destinations is also explored in Kieu et al. (2015). Zhao et al. (2007) partially validated a similar methodology verifying counts at an aggregate level and at the trip stage level. Later Munizaga et al. (2014) validated this approach using a sample of 601 random public transit riders who accepted that their smart cards be identified. Their reported mobilities were compared with the mobilities processed through the data produced by their smart cards. Results are: for boardings, 98.9% locations were right (small AVL errors due to the GPS system). 84.2% of alighting locations were successfully estimated. Route choice within the Metro system were good for 85%. These are very good results, and the article also provides a list of reasons why the alighting estimation failed. This article is very valuable to research on smart card data since previous research rarely include a validation step because there is no datasets which would allow it. Some faring system are distance based, or include a zone faring policy which requires from the rider to tap-in when he boards and to tap-out when he alights such as Singapore (SBS, 2016) or London's pay-as-you go system (England) (TfL, 2016).

Knowing alighting location (through inference methods or through tap-out data) does not mean that the activity location is known, some of the trip leg's destination can be a place to wait for a different bus line (connection trip). Knowing activity location allows the production of indicators that are demand oriented (similar as those drawn from travel surveys), such as average trip length (Trépanier et al., 2009), and to study point of interests and transfer stations (Chapleau et al., 2008). Therefore, more work is required to find activity location. Chakirov and Erath (2012) proposed and compared two approaches: a rule based approach and a modeling based approach. The methodologies were applied on AFC data from a system that demand tap-in and tap-out (the AFC system from Singapore). The rule based approach consists in studying the time interval between an alighting and the next boarding and setting a time threshold: activities lasting 6 hours to 16 hours are labeled work purpose and activities lasting 1 hour to 5 hours are labeled 'other' purpose. The rule based approach has the main

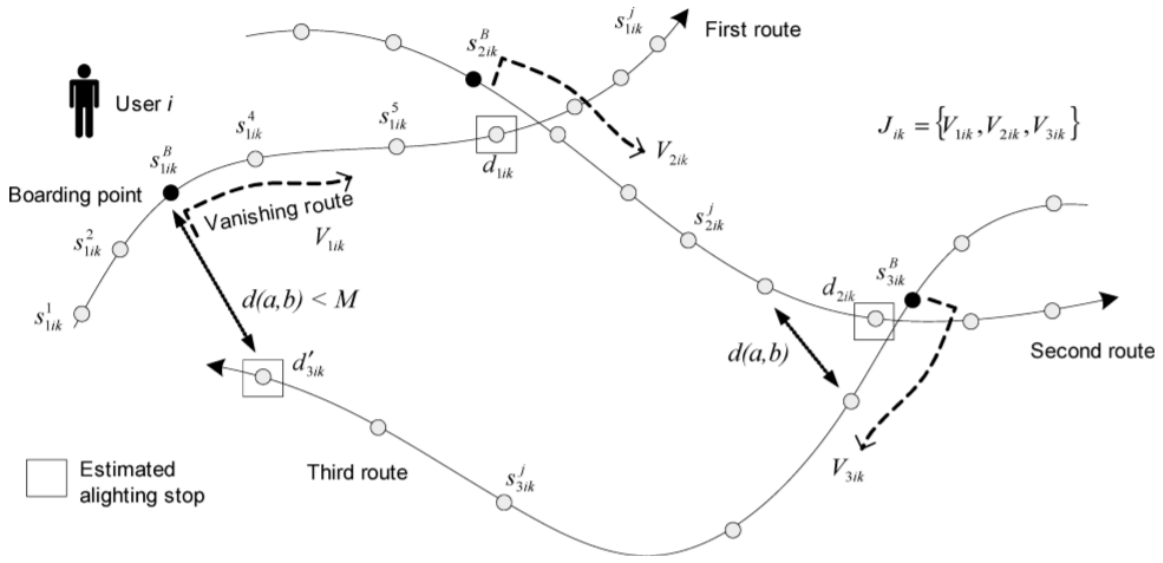

Figure 2.1 The alighting location estimation model for "normal" and "last" trips. Source: Trépanier et al. (2007)

issue of being blind to short activities and to set similar threshold for all users even though the behavior heterogeneity is known (Ortega-Tong, 2013). The rule based approach was later reproduced, adapted and validated on other case studies (Devillaine et al., 2012) and validated against reported and ground truth (Munizaga et al., 2014). On the other side, the modeling approach relies on multiple discrete choice models for activity duration, activity start time, destination's land use. It enables a more comprehensive approach, however it was demonstrated using very simple models. As noted in the article the model was calibrated on a data set including other modes than public transit trips but no mode choice modeling was done, and it did not include any socio-demographic variables which may seem as a limitation for a behavioral choice model. Some papers adopt a data fusion approach of AFC data and travel survey data relying on Bayesian methods (Kusakabe and Asakura, 2014; Zhong et al., 2014). This approach enables a great granularity of purposes however the results proved the method to be efficient only for the same kind of trip purposes than for the rule based approach (it makes a difference between long and short activities). In addition it was not validated against reported truth nor ground truth.

2.3 Expectations from AFC data

Expectations from AFC data can be categorized in three layers: strategic planning, tactical and operational (Trépanier et al., 2007). The strategic level is long term planning, since

smart card is longitudinal, it allows a better behavioral understanding of smart card users. Studies were conducted to characterize and classify smart card users. Agard et al. (2006) propose a clustering approach of the smart card data from Gatineau (Canada). They are able to identify four groups of smart card users based on departure hours of each trips stage. They also show that their clusters composition is quite stable through time even though their work do not ensure that single smart card are persistently attributed to the same cluster. It is though reasonable to think so. Ortega-Tong (2013) went way further: on London public transportation system, she considered many classic travel dimensions to characterize and classify smart card users; such as journey start time, travel distance, activity duration, fare, public transport mode choice. She also made good use of the longitudinality of the AFC data set by working with average values differentiated between week days and week ends, and by including dimensions such as travel frequency and origin-destination frequency to her clustering work. She was able to differentiate and identify up to eight clusters and the week day/ week end differentiation proved to be highly important. Clusters were validated using the intra-cluster variance analysis. She also found out that within the public transportation system, the mode choice was not homogeneous: some people would rather like staying away from the metro while other would rather like using the metro. However these clustering approaches are really limited to describe travelers. First, there are very few variables to describe clusters since AFC system does not record socio-demographic attributes other than the fare charged. Secondly the number of clusters found during the clustering process (either arbitrary defined or analytically found) is limiting the granularity available to describe the travelers. A discrete choice modeling approach will provide a better understanding of public transit travelers. Chapleau (2013) studied and characterized seasonal variability of public transit use. He compared summer and fall through demand and frequency distribution of a fare product. He found significant differences. The longitudinality of smart card data is a very important point to use since the Mobi*drive* project (Axhausen et al., 2002) proved that there is variation of behaviors over time. On the other side accomplishing a trip diary survey over many weeks is even more complex than doing a regular one-day survey since it has the same kind of biases and it includes a fatigue effect for reporting as well. By studying smart card data, more exotic results were found by doing smart card data analysis: studying Singapore, Chakirov and Erath (2011) found out that in some cases the value of the certainty to get a seat is highly important in the decision making process. All these studies give a better understanding of public transit riders and enable public transit riders to adapt their supply and their faring policy to their users. Trépanier et al. (2012) specifically studied smart card users loyalty. It is also expected that smart card can be used to control travel surveys validity. Spurr et al. (2014) point out that travel surveys mislead to consider bad demand estimation.

They compared public transit use estimated from the Montréal (Canada) travel survey to the observed public transit use (through smart card data) and found out underestimations and over estimations up to 20%. In a later paper (Spurr et al., 2015) the same authors proposed a heuristic methodology to match smart cards to their owners when they were interviewed for the travel survey. Their methodology provided results for approximately 50% of the population that were surveyed/ However, there was no validation against ground truth. Then they individually compared reported mobility and observed mobility. They ended up profiling three types of survey respondents: type 1 are ideal respondent who report their daily mobility as accurately. Type 2 are the occasional transit user making undeclared trips causing under estimation of demand. Type 3 are the respondent who reported his typical day instead of his actual trip day. However, this heuristic approach should be considered with caution: it does not provide a 100% reliable information since the reporting problem can work in both direction and lead to assign a smart card to the wrong person who misreported his mobility.

At the tactical level, smart card data is very useful information since it is linked to the transportation system at the route level (Trépanier and Chapleau, 2001a). It allows to get very detailed information such as the maximum load for each single bus route. However, it is very important to note that the reliability of the information we can get is depending on the penetration rate of the smart card and it is not able to provide information as good as passenger counters, which count smart card users, non smart card users and free riders . We already introduce all the work that has been done to enrich smart card data in system where only tap-in is available. Utsunomiya et al. (2006) show that there could be a better suitability of supply to the demand by looking at week days separately. Among the interesting topics reported in Pelletier et al. (2011) there is the study of transfers and the impact evaluation of new public transit policies. The ideal trajectory from an activity to another is the straight line. In the public transit system, a good proxy would be a trip with the fewest connections since every time the rider has to make a transfer, he has to wait for the next vehicle. Hofmann et al. (2009) study connection trip further. White et al. (2010) state that a faring policy could be applied, adjusting price to peak-hours, off-peak hours. Faring policies based on the AFC system is a way to impact the demand that has been studied (Lovrić et al., 2013; Copsey et al., 2014). An important aspect to study of smart card data that has already been pointed out several times in this paper is the smart card holder: even if there is a high penetration rate the smart card owner can decide to stop using public transportation. Public transit loyalty is an important point since it is a source of revenue and it is also a way to estimate a stable demand (Bass et al., 2011; Trépanier and Morency, 2010). It has to be noted that loyal public transit users can move their home from one neighborhood to another which affect

spatial distribution of demand, or the smart card user can lose its smart card, therefore these kind of studies can be complex.

At the operational level, smart card data is used to generate precise performance indicators such as schedule adherence and person-kilometers (Trépanier et al., 2009). It is also used to detect irregularities and errors in the AFC system and to identify defective equipment (Deakin and Kim, 2001). We will refer the reader to Pelletier et al. (2011) as it is a good and recent literature review.

We noted that the use of smart card data for behavioral analysis is limited by the lack of socio-demographic attributes available. They are mostly clustering analysis of mobility behaviors without any socio-demographic description of the clusters. Very few paper try to attach some socio-demographic information to the smart card and when they do, it is considered as a side problem with no real strategy to tackle it.

2.4 Limits to smart card studies

For both alighting and activity location inference, the rule based approach is the most documented and developed. It is also an approach which is both spatially and temporally transferable as long as thresholds are adjusted to each case study. However there are some limits to a rule based approach which motivate other strategies. Even though the rule set may be quite good for most of the population, it can fail for an entire part of the population because their mobility patterns are widely different: Bagchi and White (2005) give the example of pensioners in Southport, Merseyside, and Bradford whose activities can be shorter than the 30 min time threshold. Amongst the most frequent failures reported by Munizaga et al. (2014) we find overcrowding of buses which delay people from taking the next bus, low service frequency which result in inferring an activity when the traveler was instead waiting for the next bus, short activities impossible to detect, errors due to the time threshold to infer work and other type of activity etc. These failures are the cause of unlinked trips (with no destination inferred) and wrong destination locations. It is creating an inconsistent database.

Articles presenting work that achieved good results are conducted on transportation system benefiting from a good AFC system with a high penetration rate of their smart card, coupled with an AVL system which enables the location of transactions. However, as Erhardt (2016) states: not all transportation systems are equal. He presented the case study of Bay area (United States of America) and showed that penetration rates are not equal in various neighborhoods and in various time window (peak hour etc). He emphasizes how badly it can impact on bus boarding counts estimates and proposes a model of smart card ownership

calibrated with on-board travel surveys. With the advent of new technologies, it is reasonable to think that the penetration rates will be higher in a near future. It enhances the estimates, however results are still quite poor compared to what could be achieved in transportation systems with good penetration rates. In transportation systems with no AVL systems, a very detailed work has been done by Gaudette (2015); Gaudette et al. (2016) to attach location information using GTFS (General Transit Feed Specification) data and spatial anchor points such as metro stations. The methodology is providing good results, however it is still adding a layer of errors.

2.5 Population synthesis

Usually, an agent-based model is calibrated on a sample of the population, in order to be applied to the whole population. However in most cases the survey has comprehensive information such as social, demographic and economic information about the population sampled. This level of detail is not likely to be known for the whole population because it would be too expansive and because of privacy concerns (Müller and Axhausen, 2010). In order to get results, the model should be used on the whole population. The process of population synthesis consists in creating a synthetic population which properties are similar to the real population, so models can be applied. Classical population synthesis methods are based on either Iteration Proportional Fitting (IPF) or combinatorial optimization. Müller and Axhausen (2010) provide a good literature review of these two methods. A problematic that population synthesis is facing is the evolution of the population over the years: surveys and census are conducted every 5 to 10 years and the population is aging, moving in or out of the zone, dying. Models exist to simulate this evolution (Eluru et al., 2008), one of the most comprehensive demographic model can be found in ILUTE (Miller et al., 2011).

The IPF was first described by Deming and Stephan (1940) consists in fitting a disaggregated bunch of agent to aggregated marginals over agent attributes. This is done using contingency table. Once contingency table is fitted, the sample population is expanded with an individual factor that insure that all marginals are respected. It is not necessarily converging since if there is a zero-cell, the IPF method contains a division by 0 operation which is undefined (Pukelsheim and Simeone, 2009). Another important limit is computation cost which grows exponentially with the number of attributes being synthesized, Pritchard (2008) proposes a list-based representation of attributes to avoid including a lot of sparse value in the computation. Memory cost is a major point limitation of IPF methods (Müller and Axhausen, 2010). In its more general description it is not able to handle a complexe structure (like individual and family), however later developments have accomplished this step with the

Iterative Proportional Update (Ye et al., 2009)(Müller and Axhausen, 2010). It is the most common population synthesis method. The combinatorial optimization method is described in Voas and Williamson (2000). It has the benefit of being able to handle data with very poor spatial repartition and output data at higher spatial resolution. It uses simulated annealing to find the best composition of agent to fit marginals. Then, in the same way IPF does, it expands the sampled population with weighting factors. Ryan et al. (2009) and Huang and Williamson (2001) propose a comparative study of IPF and combinatorial optimization methods. They found that combinatorial optimization provide more stable population synthesis. It is also simpler to implement and less sensitive to sampling errors. However combinatorial optimization is sharing a major limitation with IPF: they are both cloning sampled data.

There is a recent third methodology based on Monte Carlo Markov Chains sampler (Farooq et al., 2013a) which release the marginal fitting step. Two samplers are used either jointly or combined: Metropolis Hasting and Gibbs samplers. Instead of computing weighting factors for individual, it samples agent from the joint distribution underlying the sample and the universe population. It is done by randomly choosing an attribute then redistributing it according to the joint distribution of the population (see figure 2.2). Conditional distributions of attributes are known through the datasets available. The whole process is a Monte Carlo Markov Chain. It is able to handle two spatial level of information using importance sampling and it may be able to handle more. Farooq et al. (2013b) present a case study of this methodology on Brussel. They generated households agents with eight attributes. The methodology is applied in a software SimPSinz that is made open source. Its application is compared to an IPF approach application. The SimPSinz is stated as way more fast than the IPF approach. The framework presented is able to handle incomplete information (conditional distribution that are not known over every attributes), Farooq et al. (2013c) implements the methodology and recreate missing information using models. Farooq et al. (2013c) also contains an extensive comparison between IPF population synthesis and Gibbs sampling. The Gibbs sampling performs with more stability and without significant loss of marginal fitting (always $\leq 1\%$). The performance of population synthesis was assessed for various sample size (5%, 3% and 1%). Even if simulation based population synthesis is a recent methodology, it seems to provide more reliable results. Anderson et al. (2014) proposes an association method to distribute a role within the household. The methodology is based on the Hungarian algorithm (graph theory alrogithm): one set is the population, the other set are household roles available (head, spouse etc). Edges are weighted using utilities simulated from a multinomial logit model and the Hungarian algorithm is used to minimize the cost of associating a role to every person in the population.

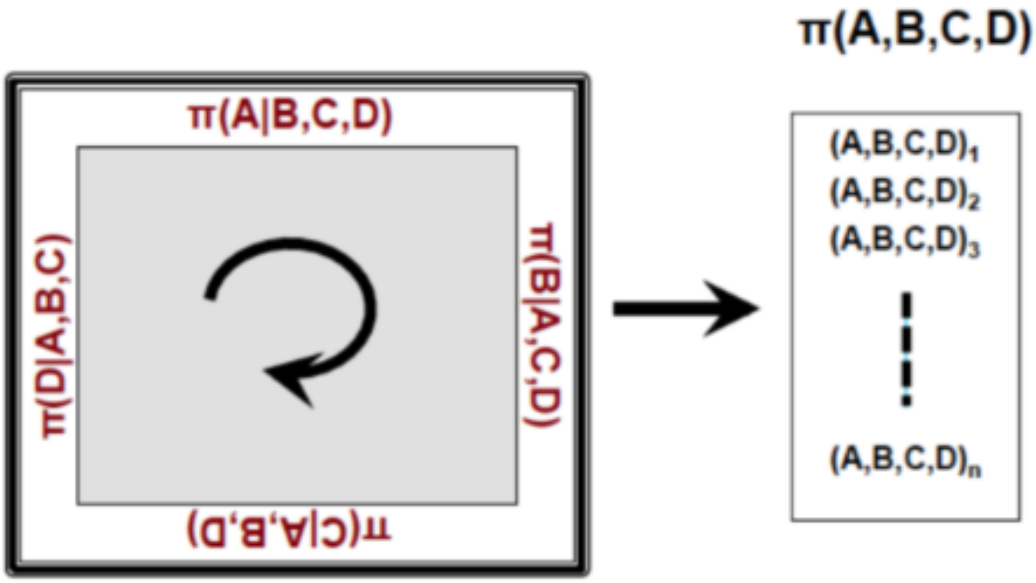

Figure 2.2 Illustration of the working of Gibbs sampler. Source: Farooq et al. (2013a)

2.6 Trip chain choice model

When enriching smart card data with rule based approach, one of the major drawback is that it is blind imputation without considering any behavioral aspect. It results in poor granularity in the outcome. The modeling approach did not use any socio-demographic variables (because there wasn't any available). Bonnel (2002) and Meyer and Miller (2001) defines the mobility of an individual, making a difference between the trip, the activity, the trip leg, the mode, the frequency of trips and activities. All this information is available or can be inferred from smart card data for public transit users, but there are few work to model the entire mobility. Some models focus on modeling long term decision reporting such as mode choice (Mokhtarian, 1991), place of living (De Palma et al., 2014), car ownership (Vovsha and Petersen, 2009) and other onetime choices with persistent effect on daily life. Roorda et al. (2009) propose a model combining activity scheduling and long term decision based on the definition of a stress function measuring actual state and distance to ideal state for each household. In a more general way, Land Use Transport Integrated models (LUTI) are good examples of work that model those interactions between long term and short term decisions: projects such as UrbanSim (Waddell, 2002), ILUTE (Salvini and Miller, 2005) (Miller et al., 2004)(Miller and Salvini, 2001), ILUMASS (Wagner and Wegener, 2007), TRESIS (Hensher and Ton, 2002) etc. There are two kinds of LUTI models: agent based and aggregated

models, agent based model are more relevant to our work since we are trying to work at the smart card holder level. See Renner et al. (2014) and Bierlaire et al. (2015) for more detailed literature reviews on LUTI models.

In a complete mobility, there are many choices to accomplish (mode choice, number of activities, activity locations, departure hours etc) and these choices may be correlated, for instance having a great number of activities in various locations may be an incentive to use a car. There are two ways to model those choices: either jointly or sequentially (Doherty and Axhausen, 1999). Sequential models are assuming that there is a natural causality between choices that can be captured and that there is a logical order of choices which creates a nested structure of choices. Most tour based model are relying on a 'tree logit' form (Miller et al., 2005) .Bowman and Ben-Akiva (2001) propose a five nests structure under the main assumption that there is a natural difference between the main activity and secondary activities and that the main activity structures secondary activity choices. This is an important assumption that was later used in other sequential models: FAMOS (Pendyala et al., 2005) TASHA(Miller and Roorda, 2003)(Roorda et al., 2005) and more generally all the activity choice models generating full day schedule in a LUTI. In models which strategy is scheduling the time use (based on Hägerstraand (1970) work), sequential models are often used. Another good point about sequential model is that they can be included in Computer Based Model (CPM) because their different layers can be applied dynamically in unlinked parts. Used this way, they can be estimated separately which is much easier. This alliance between CPM models and stochastics model is called 'weak CPM' in Yasmin et al. (2015).

One of the major downside of sequential models is that they are ignoring endogeneity that is not explicitly considered in the tree structure. On the other hand, joint models does not rely on any causal assumptions and are able to capture those latent effects (Guo and Bhat, 2001) Golob (2003). Joint models have been used for estimating pollution(Paleti et al., 2011), study latent demand for new highways (Fujii and Kitamura, 2000) Bhat and Singh (1997) propose a joint model of work mode choice and secondary stops (other than work), they are able to get a good understanding interaction between choices. Golob (2000) proposes a joint model of household activity participation and trip chain generation. Joint models present strong theoretical ground, however from a practical point of view joint models are limited and cannot consider too many attributes and attribute's categories since traditional model estimation techniques are not performing enough (Bhat, 2011). Bhat and Eluru (2010) provide an extensive literature review of recent econometric models.

CHAPTER 3 METHODOLOGY

Our methodology is independent of the datasets available, however it is more understandable if presented with a context. At first, we present the available information, then a broad description of the methodology is given. Each step of the methodology is detailed. Eventually, the ideal validation method is described.

3.1 Context

This section presents the datasets that we used.

3.1.1 Smart card dataset

The research was conducted using AFC data from Gatineau's public transit system (Canada). The public transit system operator is Société de Transport de l'Outaouais (STO). STO is serving a population of 259 800 persons (according to iTRANS Consulting Inc. (2006)) and it is covering an area of 637 km^2 (iTRANS Consulting Inc., 2006). A very detailed historical description of this organism is available in Blanchette (2009). We use data from November 2005, at that time there was a bus-based public transit network (no metro, no bus rapid transit) with 112 bus lines (figure 3.1 is a map of the location of bus stops), including a night service. For this period, there are 23 549 smart card holders which represent close to 80% of public transit users (Blanchette, 2009). The whole fleet was equipped with AFC system in 2001 as well as an AVL system therefore the data is quite reliable since there was time to find errors, correct them and improve the AFC and the AVL system. Data available contains, for each smart card transaction:

- date and time of boarding
- fare (regular, express, interzone, student, senior)
- line number and direction
- boarding bus stop number
- alighting bus stop number
- other information less relevant to our purpose

Alighting bus stops were inferred using Trépanier et al. (2007) methods. There are 831 119 validations for November 2005 and the alighting inference algorithm resulted in estimating successfully 718 539 trip leg's destinations (86%). A distance threshold of 2 km was used to detect missing trip (the walking distance threshold is 1 km, but it is assuming that there could be an activity in between the two stops). The remaining 14% is kept in our dataset since the developed methodology is not affected.

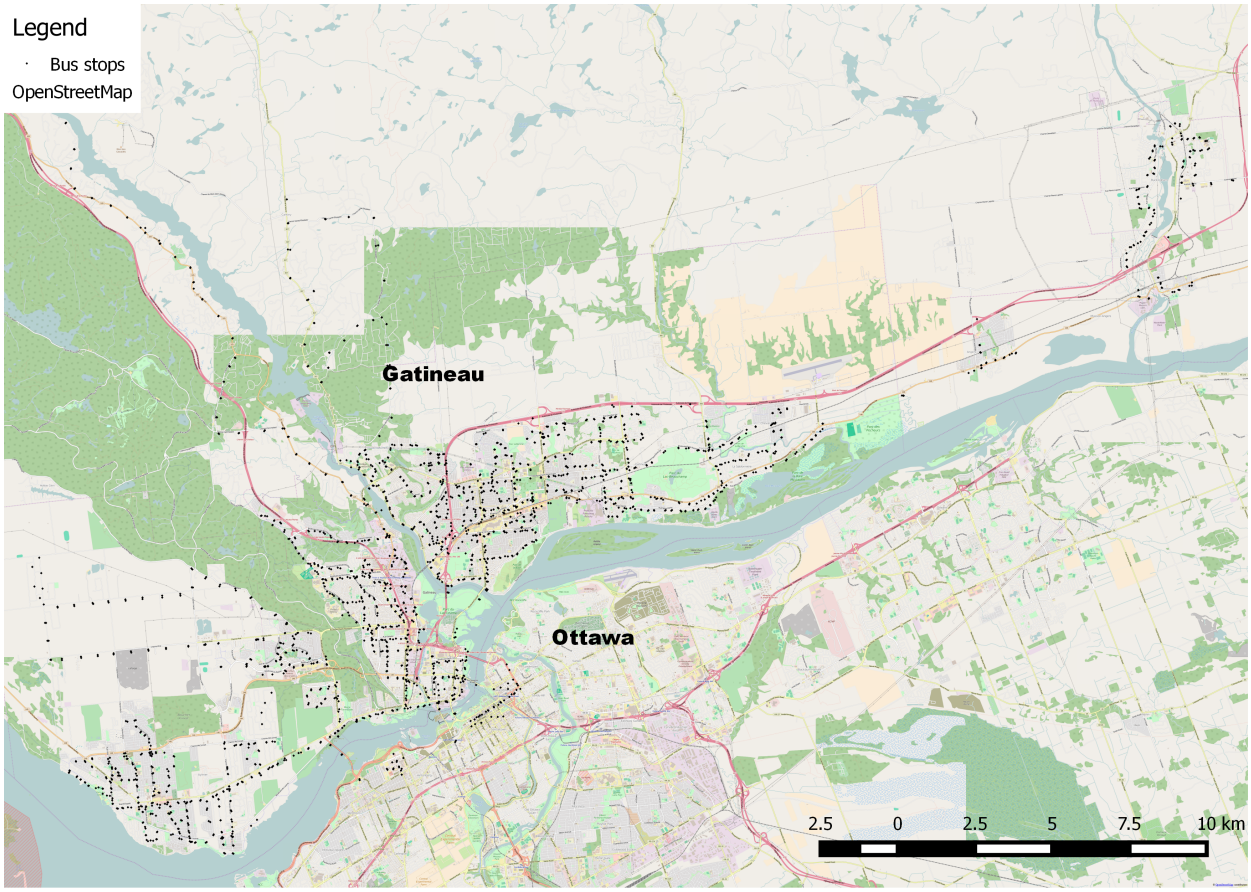

Figure 3.1 The bus stops of the STO's network.

3.1.2 Origin-Destination survey

The travel survey dataset of the metropolitan area of Ottawa-Gatineau is the Origin-Destination survey. It is a land-line based survey with a sampling rate of 5.1%. It represents an average week-day, reporting only for 11 years old or older. The whole family's mobility is reported. It is actualy a trip diary survey for the whole day for the entire family. The survey was conducted between September 21st, 2005 and November 29th, 2005. The metropolitan area

of Ottawa-Gatineau is divided in two by a river (see figure 3.1). The STO is running a bus service in the Gatineau area and part of Ottawa downtown. In this area, public transit trips represent approximately 20% of all trips (iTRANS Consulting Inc., 2006). However public transit trips include trips on the STO network as well as trips on the school transportation network or other public transit services (such OC transport, the Ottawa public transportation network). Only 13.5% of all trips in the Gatineau area are made on the STO network. Therefore, with the smart card data we have approximately 11% of all trips made on the STO network in the Gatineau area.

3.1.3 Canadian census

The Canadian census dataset is ran by Statistics Canada and provides good information about the whole population. It provides marginal distributions of attributes such as age and gender at the dissemination area level. Dissemination areas are spatial areas which are designed to be uniform with a population count targeted between 400 and 700 individuals (Canada, 2016), therefore, in dense urban neighborhood, this can result in a zone with a area of a few hectares (15 to 50 ha). It was held on May, 2006.

3.1.4 Public Use Microdata Sample

Public Use Microdata Sample (PUMS) is an extension of the Canadian census. It samples 20% of the whole population. It is a disaggregated sample, therefore cross-distribution information can be drawn. For privacy concerns geographic information is limited to a very large scale (at the metropolitan area level). It contains information about age, gender, family size, income, language etc. It was held on May, 2006.

3.2 The broad picture

Our data evolves in a complex space that we reduce to four dimension for explanatory purpose. First, the spatial dimension. A strong spatial information is known at a very low scale. Secondly, temporal dimension. A strong temporal information is available at a disaggregate level for a long time window (longitudinal data) whereas a weak temporal information is a one day information. Third, the socio-demographic attributes: a strong information is known at the disaggregate level while a weak information are marginal distributions. Fourth, the mobility choice dimension. A strong information is the knowledge of the complete trip chain including the mode use, time of departures, activity purposes etc (see figure 3.2) while a weaker information would be missing something.

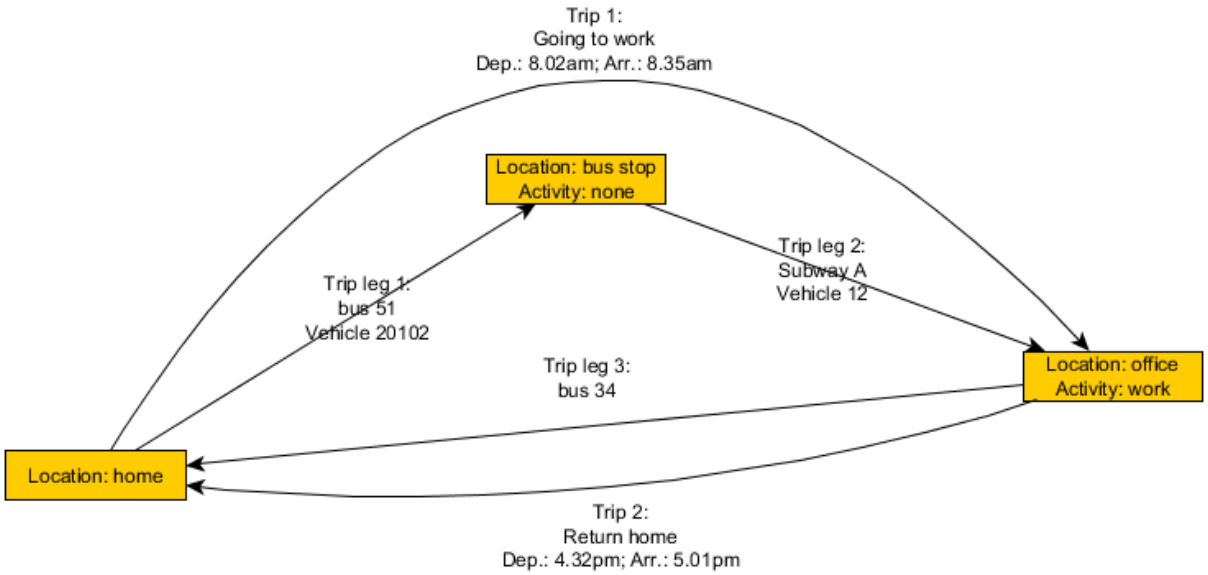

Figure 3.2 Decomposition of a trip chain.

Table 3.1 Problem dimensions.

Dimension	Spatial	Temporal	Socio-demographic	Mobility choices
Travel Survey	Weak	...	Medium	Strong
Census	Strong	...	Medium	...
PUMS	...	...	Strong	...
AFC data	Strong	Strong	Weak	Medium

To make the best out of our data strengths and weaknesses (see table 3.1) we came up with three core hypothesis. The first hypothesis is that the first boarding of a user is within the neighborhood of the houser's home (this is derived from the assumption that there is a walking distance threshold). Therefore, if we know the most frequent daily first boarding, we know the smart card holder is to be found in the local population which can be known through census. The second hypothesis is that individuals have various travel behaviours with respect to their socio-demographic attributes. Therefore, within the local population some persons are more likely to be the owner of a smart card that fit their travel habits. Travel habits can be modeled. The third hypothesis is that the faring policy is well understood by everyone, it induces that the fare applied can provide some information about the smart card holder's attributes (there is often a social faring policy with student fare, retired fare etc). We expect that those combined assumptions can allow us to attach socio-demographic attributes to smart cards. Figure 3.3 illustrates the process.

The first stage of our methodology is to use our first hypothesis: for a given spatial zone, we have the local population and smart cards that are based in the spatial zone. On the first hand, the population is known through the Canadian census data, the PUMS data and the travel survey and its information can be known at the dissemination area level (smallest geographic area available for census data: it is designed to contain approximately 700 persons). On the other hand, we assume that the most frequent station for daily first boarding is a good proxy to estimate smart card owner's living area. This gives us a set of persons and a set of smart cards which are from the same spatial zone.

The second stage consists in weighting possible links between the two sets using our second hypothesis. From a smart card records, we observe a travel pattern on the public transit network. We use behavioural discrete choice model: each individual in the population is an agent who can choose a travel pattern among the smart card set or choose not to use public transit. The discrete choice model provides us with utility functions and probabilities which can be used to weight the links between the population and the smart cards (a similar methodology was used in Anderson et al. (2014) to assign individual agent to a household role). The idea to work the other way around (let the smart card choose its owner) was raised. However socio-demographic characteristics are not only correlated to travel patterns, they also have a causal effect on them. As it is a causal link, we can not make the hypothesis that knowing the mobility of an individual is enough to infer his attributes. There is no bijection between socio-demographic attributes and mobility patterns; as public transit users clustering work has shown, different types of individuals can have a similar travel pattern (Ortega-Tong, 2013).

The third stage consists in actually assigning a single smart card to a single owner, assuring that this is the most probable solution. This is done using the Hungarian algorithm (Anderson et al. (2014) used it in a similar way). This solution was chosen because it was the most straightforward, but any association algorithm can be used.

The implementation and the work flow of this methodology is illustrated in figure 3.4.

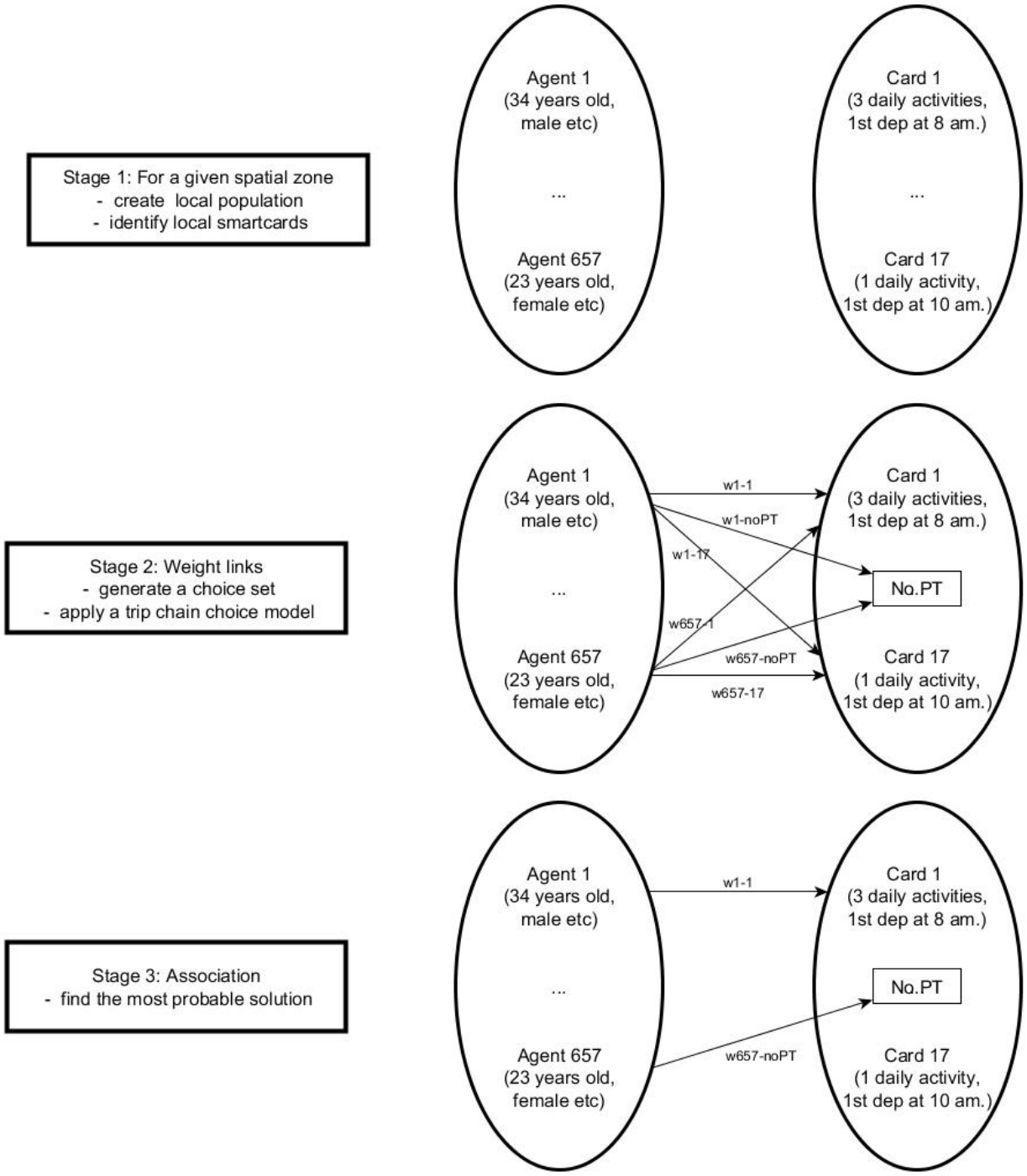

Figure 3.3 The three stages of the methodological framework.

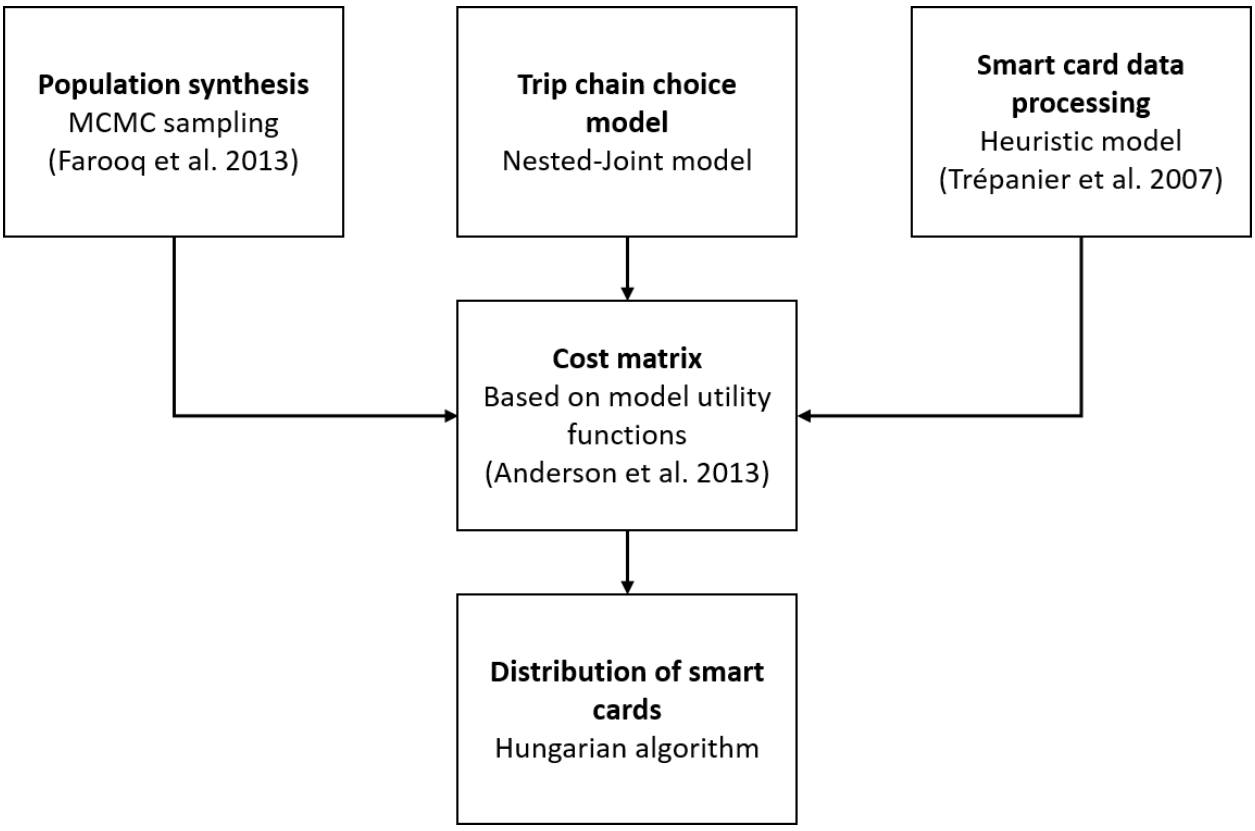

Figure 3.4 Work flow of the methodology.

3.3 Enriching smart card data to the trip chain level

Smart card data is enriched with trip leg's destination using rule based approach (Trépanier et al., 2007). Figure 1 summarizes the heuristic. We make the assumption that there is a walking distance threshold above which travelers will not walk (we set this threshold to 1 km, according to Egu (2015) analysis and Trépanier et al. (2007) example). For a single user on board of a bus, we call $R = s^j$ the route of the bus, where s^j are the route stops. The vanishing route (Trépanier et al., 2007) is the sequence of stops where the user can alight after his boarding. We define the distance between the vanishing route V and the next one W as:

$$d_{V \to W} = \min_{s^j \in V}(dist(B_W, s^j))$$

B_V is the boarding stop, $V = s^j, \forall j \geq B_V$ is the associated vanishing route. We note $W = s^j, \forall j \geq B_W$ the next vanishing route.

A distance threshold of 2 km is set to infer a missing trip between two boardings (it is twice the walking distance threshold since there could be an activity in between). If $d_{V \to W} \leq 2km$ we assume the alighting station is $\min_{s^j \in V}(dist(B_W, s^j))$. This model estimates 'the best location'. The case of last trip of the day is special since there is no following vanishing route. We treat the last trip of the day by buckling it with the first trip of the day, or if it fails with the first trip of the next day. The assumption being that most people come back home at the end of the day and take public transit close to place where they spent the night. Some alightings may still be unknown. We make the assumption that travelers have a regular behavior: if a boarding can not be linked to a following one (because there isn't any), we look at historical data. If in historical data there is a trip leg with the same vanishing route that started at approximately the same time and whose alighting is known, then we assume that they both have the same alighting. In the end, there are three reasons for not being successful to infer an alighting: a) data not valid (errors cannot be corrected), b) too far (distances between locations are greater than the threshold and no regularity analysis can be done, c) trip are single and regularity analysis cannot be done. The sudo 1 describe this methodology with more details.

Data:
smart card data: boardings $((s^j)_k)$, for $k = 1..N_{smartcards}$ & $j = 1..J_k$
bus routes: $(R_l) = ((s^j)_l)$, for $l = 1..N_{routes}$ & $j = 1..J_l$
stops geographies: (s^j), for $j = 1..N_{stops}$

Result:
Inferred alighting stations

```
initialize the oriented object model;
set walking distance threshold: d_w;
for sm ∈ Smartcards & t ∈ dates do
    create daily vanishing routes: V = (v^j)_{sm,t};
    for v ∈ V do
        if v is last of day & v is not first trip of day then
            w = V.getFirstVanishingRouteOfDay(t);
        else if v is last of day & v is first trip of day then
            if V_{sm,t+1} is not empty then
                w = V.getFirstVanishingRouteOfDay(t + 1);
            else
                no destination can be inferred for this boarding;
        else if v is not last of day then
            w = V.getNextVanishingRoute();
        end
        d =dist(v → w);
        if d ≤ d_w then
            s such as s ∈ v & dist(s → w) = d;
            s is the alighting stop of v;
        else if ∃ v_{t−1} such as v_{t−1} ≈ v then
            if v_{t−1} alighting was inferred then
                s = v_{t−1}.getAlightingStop();
                s is the alighting stop of v;
            else
                no destination can be inferred;
            end
        end
    end
end
return smart card data with alighting stops;
```

Algorithm 1: Destination inference heuristic.

Activity locations were inferred using the rule based approach (Devillaine et al., 2012). In our framework, the activity location is the closest stop to the actual activity location. We

decided not to consider activity type since it induces more heuristic work, including more systemic errors and the outcome can't have a high granularity (the higher the granularity, the higher the errors). However an activity type could be an outcome of our method. A time threshold of 30 min between two successive boarding is set to infer an activity location (this time threshold is based on previous studies on the Gatineau area (Devillaine et al., 2012)). The impact of these values can be explored by making this parameter vary (Egu, 2015). It has to be noted that it would more relevant to use a time threshold from the time of alighting to the time of next boarding. This could be done using routes data more intensively.

3.4 Population synthesis

The population synthesis is the first step of the methodology. It consists in creating synthetic agents to which we can apply the trip chain choice model. As it was presented in the literature review, the Gibbs sampling method (Farooq et al., 2013c) (Farooq et al., 2013a) (Farooq et al., 2013b) seems to perform better and the software is open source. It is a Monte-Carlo Markov Chain (MCMC). The final goal of this work is to enrich smart card data with socio-economic attributes through a trip chain choice model, therefore we have to synthesize attributes that will both bring analytic power to the smart card dataset and provide explanatory power for the trip chain choice model. We also have to consider that we should limit the number of attributes for two reasons. First it is making the computation cost increase substantially. And secondly, joint distribution of attributes is known through survey data, synthesizing a population with a high granularity of attributes and high granularity of categorization of attributes results in distributions that are not consistent because of a lack of data. Even though 5% is a high sampling rate, when trying to sample from the underlying joint distribution, the information can be highly scattered because of numerous attributes. The methodology for population synthesis with Gibbs sampling is detailed in Farooq et al. (2013c), we will summarize it for the reader.

We consider the true population constituted of agents characterized by a set of n attributes $X = (X^1, X^2, ..., X^n)$. Their unique joint distribution is noted $\pi_X(x)$. This distribution is approximately known since it can be observed through surveys: $\pi_{\hat{X}}(\hat{x})$. The goal of a population synthesis is to draw a synthetic population which distribution is as close to $\pi_X(x)$ as possible. The Gibbs sampling is a technique that allows to do that. It requires to know all conditional distributions of each attribute over the other attributes (every $\pi(X^i|X^j = x^j$, for $j = 1, ..., n \ \& \ i \neq j)$) or at least approximations (Farooq et al. (2013c) propose either to simulate missing attributes or to assume that the distribution over the unknown dimension is uniform). As it is described in algorithm 2, the Gibbs sampling process needs to be initialized.

This can be done by taking a data point from the data available, as a seed to inseminate the sampler. The Gibbs sampler needs a warm up (Geyer, 1992), which is basically running the Gibbs sampler without sampling any agents. It is required to insure that the MCMC walk is far away from the initial seed and that it is pure random walk non driven by initial conditions. A single step of the Gibbs sampler consists in randomly choosing an attribute and assign a new value to it by doing an antithetic draw from its joint distribution. Between each sample of agent we need to let a good number of steps of the Gibbs sample to insure that successive agent draws are independent. The Gibbs sampler is a fast MCMC based sampling since there is no rejection, the interval size is defined and then an agent is sampled every interval.

Data:
global distributions: $\pi(A^i|A^j = a^j \text{ for } j = 1...k \ \& \ i \neq j),\ i = 1, ..., k$
local distributions: $\pi_x(A^i|A^j = a^j, \text{ for } j = 1...k \ \& \ i \neq j),\ i = 1, ..., k$
iterations (integer): Size of the population pool
interval (integer): Acceptance interval

Result:
Draws from $\pi(x)$

---

initialize X_{prev};
initialize X_pool;
initialize counter;
for $size\_pool \times interval$ **do**
  Generate a random number from $r = U(1, k)$;
  if $\exists \pi_x(A^r|A^j = a^j \ for \ j = 1...k \ \& \ r \neq j)$ **then**
    Generate a^r_{curr} using **Inverse Transform** on
    $\pi_x(A^r_{curr}|A^j = a^j_{prev}, \text{ for } j = 1...n \ \& \ r \neq j)$;
  end
  else
    Generate a^r_{curr} using **Inverse Transform** on
    $\pi(A^r_{curr}|A^j = a^j_{prev}, \text{ for } j = 1...n \& r \neq j)$;
  end
  $X_{curr} = X_{prev}$ with x^r_{prev} replaced by x^r_{curr};
  if *counter equals interval* **then**
    X_pool.Add(X_{curr});
  end
  $X_{prev} = X_{curr}$;
end

Algorithm 2: Gibbs Sampling using local distributions when possible. This is following Farooq et al. (2013b) methodology.

One of the major points to apply this methodology is preparing the conditional distributions from every dataset available. As we described in the section 3.1 we have access to three different datasets: aggregates over age and gender from census and partial joint distributions from PUMS and the OD travel survey. It may occur that we have several distributions for the same attribute, for instance age is known through census, PUMS, and OD survey (see table 3.3). They have different spatial granularities. The spatial granularity of OD survey gives a good spatial information when it comes to draw the portrait of the population's socio-demographic composition. However, when it comes to include mobility choice, the high dimension of the space reduces the value of the spatial information. It still can be trusted at the census metropolitan area (CMA) (which is also the PUMS spatial level of information) which is very low level of accuracy since it contains more than a billion persons over an area of 5 126 km^2 (iTRANS Consulting Inc., 2006) (since the CMA area is including both the Gatineau area and the Ottawa area). Census has a high spatial granularity since distributions are known at dissemination area (DA) (they have already been introduced: $\approx$ 700person by DA). In the Gibbs sampler, we can add a layer of Importance sampling (Farooq et al., 2013b) to consider both local and global distributions. In our case, the most precious information is spatial positioning of the information. However, locally the distributions can be very different from the global distribution over the CMA. Therefore, importance sampling may not be able to find a potential agent that fits local and global distributions. That is why we decided to run a simple Gibbs sampling using local distributions whenever it was possible. You can find a list of attributes that are synthesized in table 3.2 and their respective sources for their distribution in table 3.3 (distributions that we used for our population synthesis are labeled with a 'X'). When an attribute's distribution is known through different datasets, then we apply the following rule. a) The dataset with the most accurate spatial information is preferred because our method relies heavily on matching people with smart cards in a common neighborhood therefore spatial information is the most valuable. b) If two datasets have the same spatial accuracy, then the dataset which contains the highest cross-information is preferred (for instance: $p(age|sex, nPers, mStat)$ is preferred to $p(age|sex)$). The more cross-information is available, the more consistent people's attributes will be.

Table 3.2 Description of attributes to synthesize

Attribute	**Category**	**Index**
Age group	11 - 19	0
age	20 - 24	1
	25 - 34	2
	35 - 44	3
	45 - 54	4
	55 - 64	5
	over 65	6
Gender	Male	0
sex	Female	1
Household size	1	0
nPers	2	1
	3	2
	4	3
	4 +	4
Marital status	Married or living with common law partner	0
mStat	Single never married	1
	Single other	2
Individual yearly income	<25 000	0
inc	25 000 to 50 0000	1
(in CAN$)	50 000 to 75 000	2
	75 000 to 100 000	3
	100 000 to 150 000	4
	>150 000	5
Car	0	0
car	1	1
	2	2
	2 +	3
Occupation	Worker	0
occ	Student	1
	Retiree	2
	Homemaker	3
Education	No certificate/degre	0
edu	High school diploma	1
	Apprenticeship	2
	College or other certificate	3
	Bachelor	4
	Above bachelor	5
Location	Dissemination area level	
loc		

Table 3.3 dataset source for attributes distribution (DA: dissemination area, CMA: census metropolitan area)

Attribute	**Known distribution**	**Source**	**Focus**	
Age	$p(age\|sex, loc)$	Census 2006	DA	X
	$p(age\|sex, mStat, nPers, inc, edu)$	PUMS 2006	CMA	
	$p(age\|sex, car, nPers, occ)$	OD 2005	CMA	
Gender	$p(sex\|age, loc)$	Census 2006	DA	X
	$p(sex\|age, mStat, nPers, inc, edu)$	PUMS 2006	CMA	
	$p(sex\|age, car, nPers, occ)$	OD 2005	CMA	
Marital status	$p(mStat\|age, sex, nPers, inc, edu)$	Census 2006 PUMS 2006 OD 2005	CMA	X
Household size	$p(nPers\|age, sex, mStat, inc, edu)$	Census 2006 PUMS 2006	CMA	X
	$p(nPers\|age, sex, car, occ)$	OD 2005	DA	
Income	$p(inc\|age, sex, nPers, mStat, edu)$	Census 2006 PUMS 2006 OD 2005	CMA	X
Number of car	$p(car\|age, sex, nPers, occ)$	Census 2006 PUMS 2006 OD 2005	CMA	X
Occupation	$p(occ\|age, sex, nPers, car)$	Census 2005 PUMS2006 OD 2005	CMA	X
Education	$p(edu\|age, sex, nPers, mStat, inc)$	Census 2006 PUMS 2006 OD 2005	CMA	X

As all datasets were collected within a small time window (from September 2005 to May 2006), we are making the assumption they were made at the same time. Furthermore there is no need to include a model to simulate the evolution of the population through time. We are mainly interested in distributions of attributes, and to the best of our knowledge there was no extraordinary event that may have substantially impacted Gatineau's population. It is quite safe to assume that even if the population may have changed a little, changes in attributes distributions are even lower. The categorization of attributes was made so it fits with the three datasets.

As described later, the population synthesis is a computationally demanding process and we developed a code able to handle multithreading to speed up computations.

A better way to do it would be to take every distributions from the travel survey since it will also be used to calibrate the trip chain model. Gatineau OD survey lack of some information (education level, income level, marital status) therefore it needed to be enriched.

3.5 Trip chain choice model

A behavioral choice model has four aspects to define carefully: the decision maker (the agent), the decision object, the decision mechanism and the choice set available to the decision maker. In our case, the decision maker is the synthetic agent generated by the population synthesis step. It has some basic characteristics (age, gender, number of person in household, marital status, education level, annual personal income level, number of car in the household, occupation) which may be used as endogenous variable in the utility functions.

The decision object is a trip chain. Its attributes can be estimated from smart card data. Most smart card data points are constituted of (or can be enriched with) a unique identifier (anonymized for privacy purposes), time stamp of the transaction, applied fare, location of the transaction (which vehicle and which bus stop). They can be enriched with trip leg's destination and activity location thanks to the methods already described in the literature review. Ortega-Tong (2013) makes an extensive description of observed trip chains through smart card data. We considered departure hours, mode choice and the fact that there is a big difference of usage between week-ends and week-days. From the smart card data, we compute statistics over available week-days (Mondays, to Fridays)(see table 3.4 for more detailled description):

- average daily first departure hour
- average daily last departure hour

- average fidelity to STO services (used STO for some trips, only used STO)
- average number of daily activities

Averaging over all the available week-days is consistent with Ortega-Tong (2013). Deviations from average behaviours and similar values for week-ends would be a way to improve our method. There is a temporal twist here: a)the travel survey actually gives information over a single day for each surveyed individual and this sums up to an average week day for the transportation system; while b) the smart card data is aggregated to an average day for each individual. We could come up with a more consistent way of using the historical data in the smart card dataset.

This list could be even more descriptive by sequencing characteristics into each day of the week (to reduce variability of indicators over week-days) and adding more longitudinal descriptive variables such as standard deviation of departure hours. Therefore the trip chain choice object could be very detailed, however most of the travel surveys are snapshots and longitudinal attributes can't be modeled. If we are ready to support the assumption of spatial transferability of behaviors from another longitudinal travel survey, then this lack of longitudinal data for calibration could be overcome. Joint models complexity increases a lot for every new choice attribute that is being considered and when it comes to calibrating the model over a travel survey dataset, it may occur that the data is too scarce (or the model too demanding). In our case we decided to take our choice object as a trip chain constituted of four attributes, week-day average first departure hour, week-day average last departure hour, average number of daily activities, average fidelity to public transit. Categorization of these attributes can be found in table 3.4. Peak hours were defined by looking at distributions of first and last departure hours (figure 3.6). We define public transit fidelity as $\frac{n_{PT}}{n_{nonPT}}$. Where n_{PT} is the number of trip legs done using public transit, it is known from smart card data. n_{nonPT} is the number of trip legs not using public transit. It is unknown from the smart card dataset, however, the distance threshold we defined to link smart card data in the destination inference part allows us to get an estimator $\tilde{n}_{nonPT}$. If $\frac{n_{PT}}{n_{nonPT}} \leq 0.95$ we label the smart card as partial public transit user. If $\frac{n_{PT}}{n_{nonPT}} \geq 0.95$ we label the smart card as loyal public transit user. This makes a total of 108 different combinations for agents. In practice, since we are using smart card data, we don't have access to any information when people did not use public transit. This means that we cannot differentiate trip chains with 0 activities or 0 trip legs made on public transit survey. We define a special trip chain choice as 'non public transit user' for all those. This reduces the number of available alternatives to 54 alternatives + 4 alternatives other than STO user.

Table 3.4 Description of choice attributes

Attributes	**Description**	
First departure hour	Before morning peak hour (before 7 a.m.)	0
	During morning peak hour (from 7 a.m. to 9 a.m.)	1
	After morning peak hour (after 9 a.m.)	2
Last departure hour	Before evening peak hour (before 3.30 p.m.)	0
	During evening peak hour (from 3.30 p.m to 6 p.m.)	1
	After evening peak hour (after 6 p.m.)	2
Public transit fidelity	Did not use public transit	0
	Partial public transit user	1
	Loyal public transit user	2
Number of daily activities	0	0
	1	1
	2	2
	3 or more	3

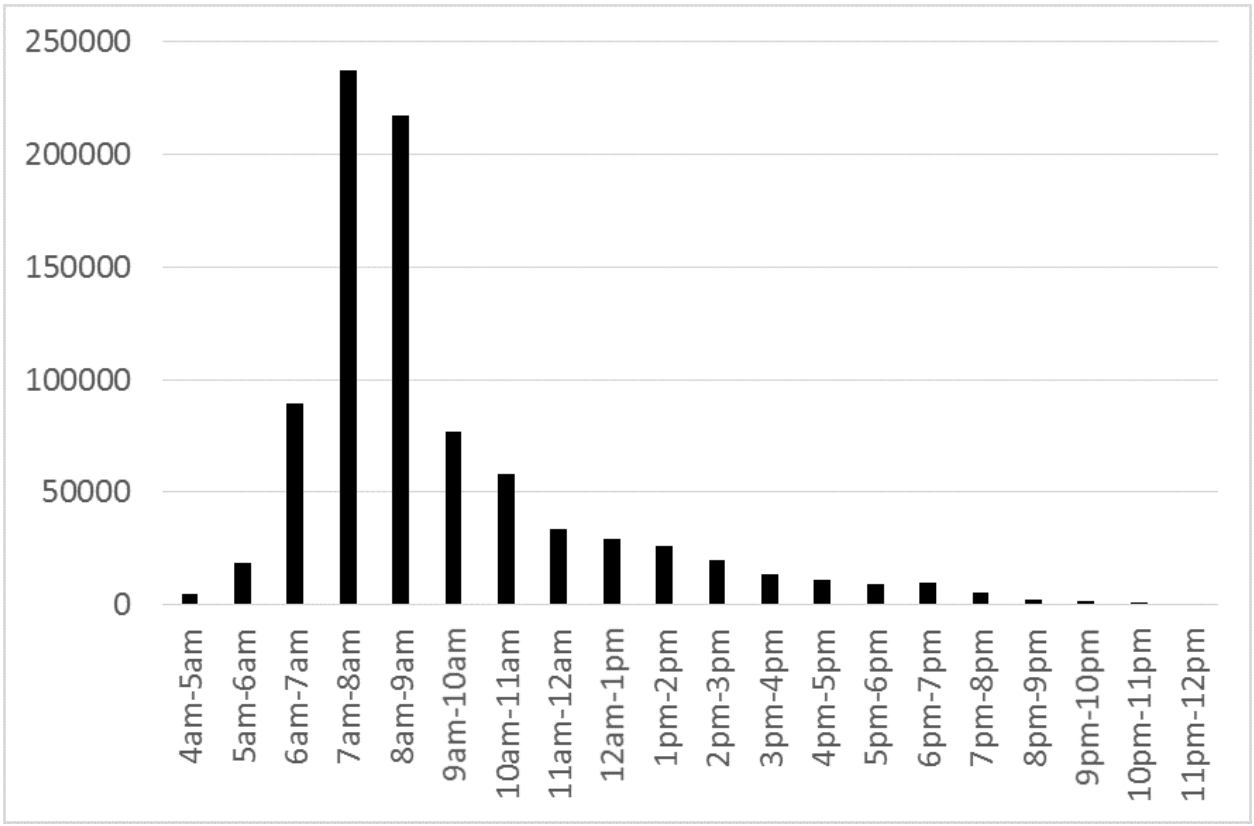

Figure 3.5 Number of first departure hours observed in the OD survey.

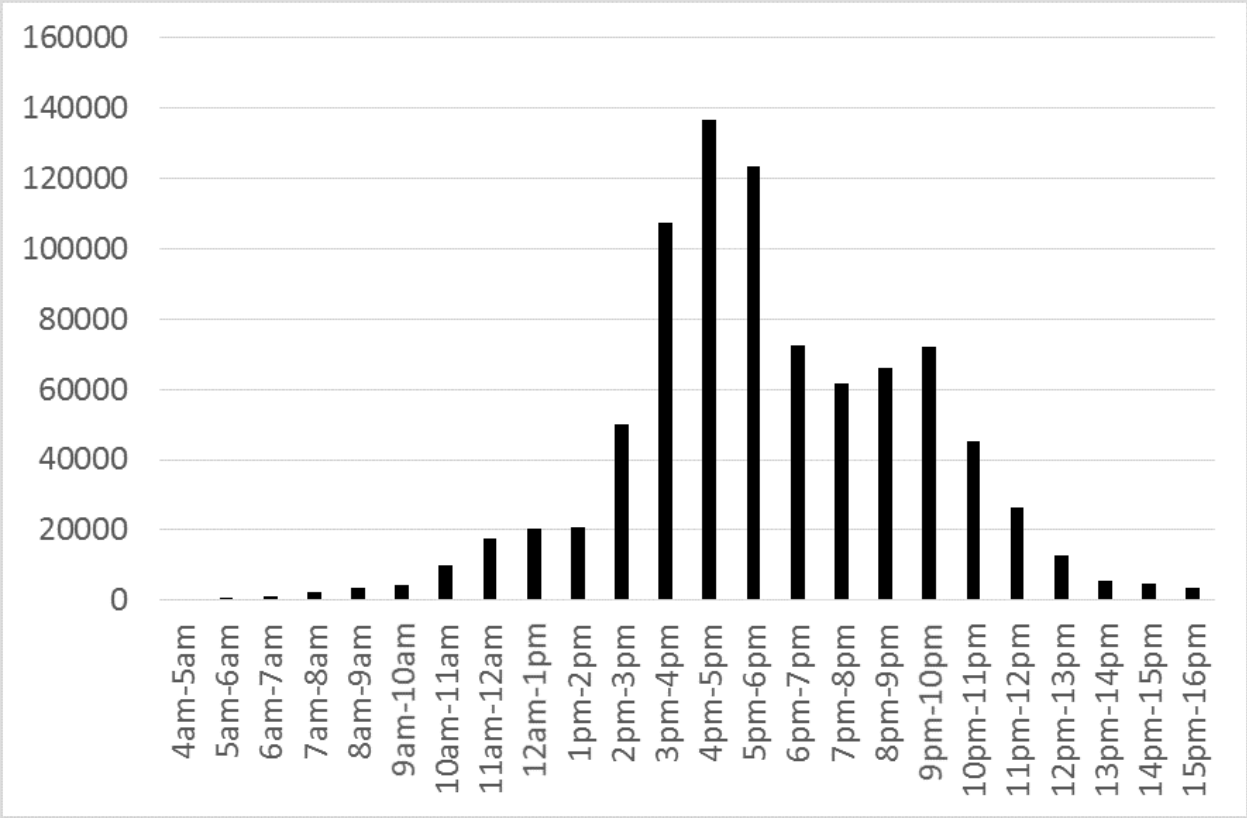

Figure 3.6 Distribution of last departure hours observed in the OD survey.

As we have seen in the literature review, there are many decision mechanisms and most of applied activity choice models are relying on a tree form because it is less demanding when generating the choice set, it is more practical to handle nests than a global architecture and because it is usable in a dynamical way (schedule based model). Joint models are less relying on causal assumptions. Hess et al. (2012) applied various model structures to the combined choice of fuel consumption and car type. The joint model estimated as a multinomial logit was found to perform as well as more complex structures (nested and cross nested structures). In our case, we have observed travel patterns and unobserved travel patterns (trips not using the STO network). We want that a) our model is able to make the difference between STO users and non STO users and b) produces results that are consistent with an observed trip chain. We propose a nested joint model as described in figure 3.7. There are two reasons for the nest structure: a) the public transit modal share was very low in Gatineau in 2005 and a nested structure was required to be able to zoom in to the level of trip chains within the STO nest and observe a difference between trip chain choices. b) the nested structure allows a considerable improvement in terms of computation efficiency by first segregating the public transit users from non public transit users and then associating a smart card to a person only on the public transit user population. For complexity reasons, we implemented a simpler structure described in figure 3.8 which shares the same nested-joint structure feature but with lower level of nests. Each nest of the upper level do not represent an actual mode choice. It represents the main mode used for the trip chain. And the trip chain always belong to the STO user nest if at least on trip was accomplished on the STO network. It has to be noted that the main goal of this structure is to segregate STO users from non STO users. We do not handle complex trip chain which do not include STO. And within the STO nest, complex trip chain are handled using the public transit usage attribute of the trip chain (only public transit or partial). This satisfies the very basic requirement for our work, however it could be developed further to the benefit of the outcome.

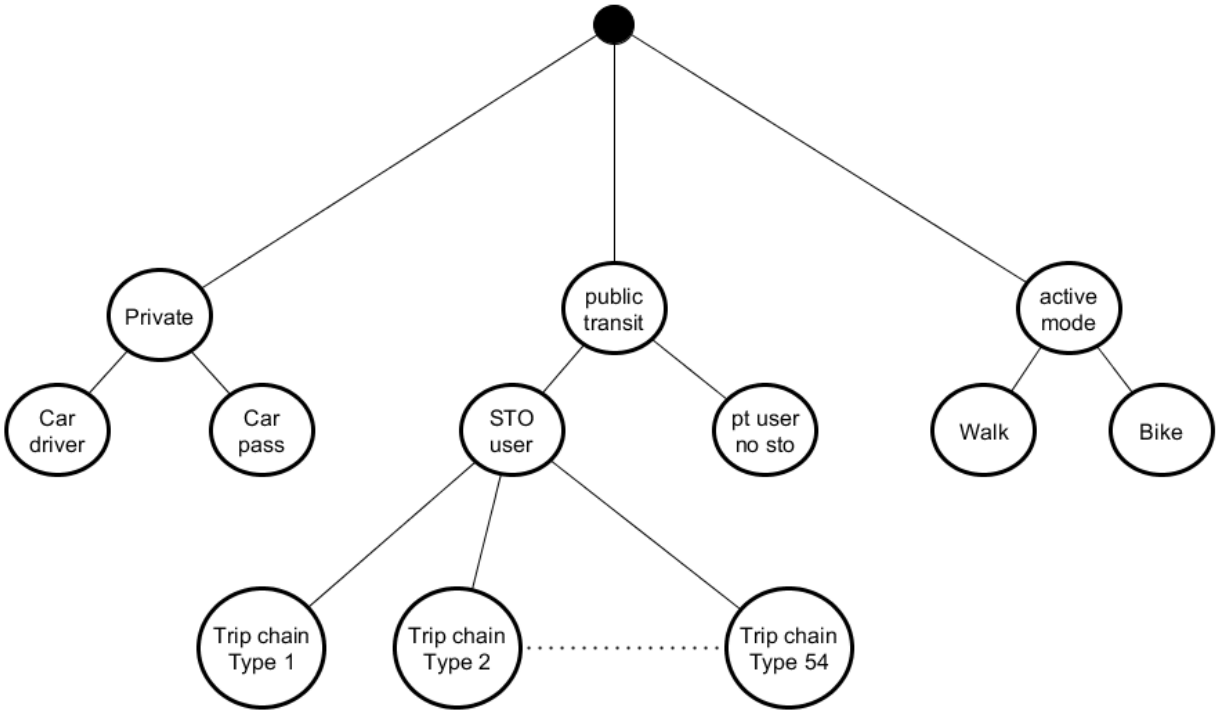

Figure 3.7 Nested joint model structure.

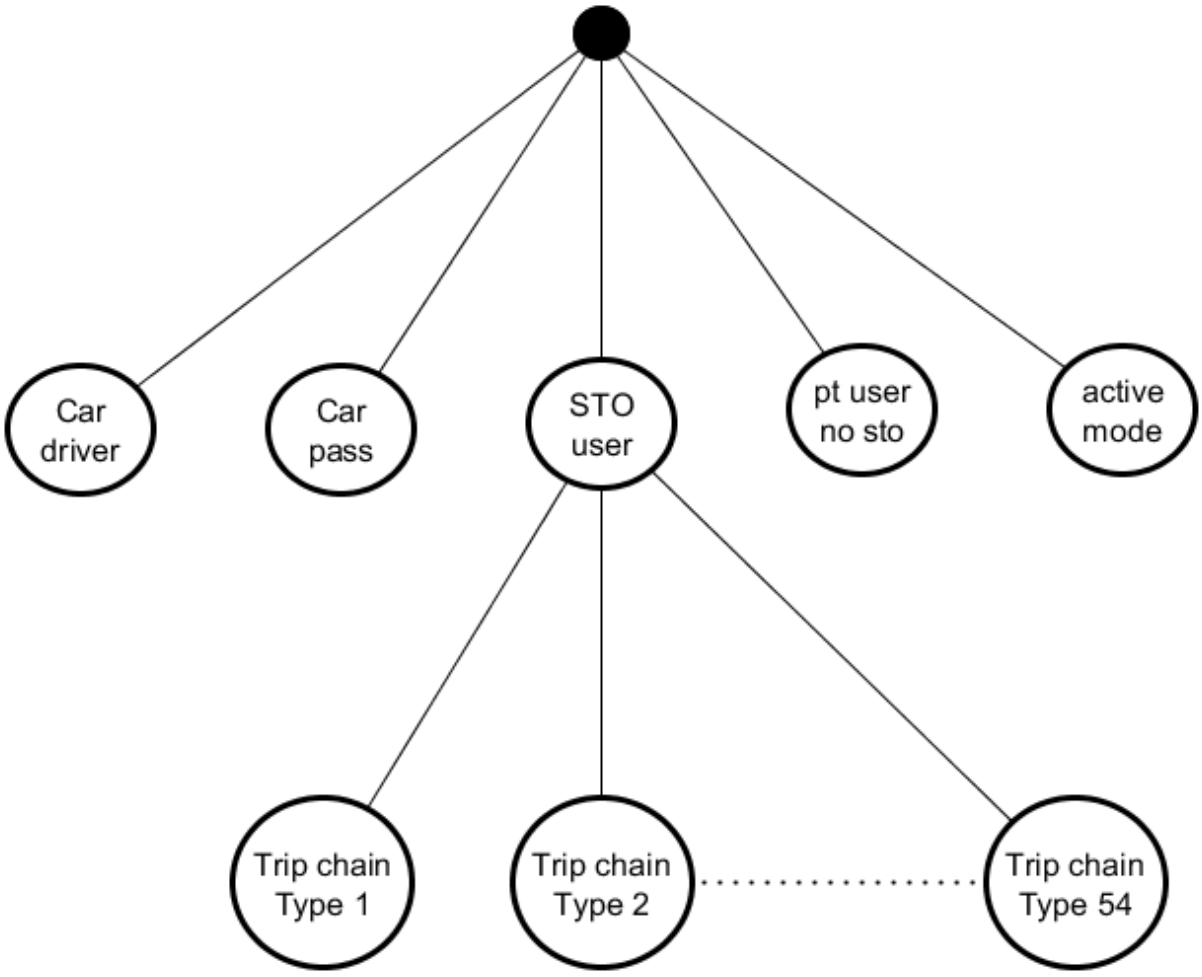

Figure 3.8 Implemented nested joint model structure.

The question of choice set generation is delicate. We are creating a trip chain choice model with 58 alternatives which is a big number therefore we may require a choice set generation strategy. There are two main uses to the trip chain choice model: in a first use, we differentiate STO users from other users (this is the first layer of nest in figure 3.8). For the second use we only need utility function to weight the links (see figure 3.3). We calibrate our model on 80% of the travel survey data and then validate it on 100%. We have two different points to check: a)we need to validate that we are able to reproduce well the population that goes into each nest and b) in the STO nest, we have to insure that trip chain simulated choices are not too bad compared to observed choices.

For the validation step (using travel survey data), the whole universe is available (58 choices). For applying the model to smart card data, the choice set is constituted of the nest choices and observed trip chain types within smart card data.

3.6 Solving the association problem

This part corresponds to stage 2 and 3 of our methodology (figure 3.3). We create and weight the links between smart cards using our hypothesis and we actually distribute smart cards using the Hungarian algorithm. For this part, we will assume that we have access to the whole population and to its spatial distribution. We also have access to the smart card data processed to the trip chain level and we have calibrated a trip chain choice model.

The actual implementation is described in the algorithm 3. The implementation relied heavily on the oriented object approach (as described in Trépanier and Chapleau (2001b) and Trépanier and Chapleau (2001a)) to alleviate computation requirements). In the first part, smart card data is processed to compute attribute values of each trip chain (see figure 3.4 for choice attributes description):

- average first departure hour during week-day (before, during and after morning peak hour)
- average last departure hour during week-day (before, during and after evening peak hour)
- average daily number of activities during week-day
- usage of public transit (not public transit user, partial public transit user and full public transit user).

First of all, we constructed the daily trip chains for each smart card data, meaning that we computed the trip chain statistics we decided to use for our behavioral choice model. For each smart card we identify daily first boarding stops. The most frequent one is labeled as the home stop. For each bus stop we identify all zones within a walkable distance and get their population (from the synthetic population): its the local population around the bus stop. We run the Hungarian algorithm on this local population and the smart cards that are associated to this stop. Once this is done, we remove every person in the local population that has been assigned to a smart card. The rest of the population is released and could be assigned to a smart card for another bus stop.

Data:
monthly smart card data: boardings $((s^j)_k)$, for $k = 1..N_{smartcards}$ & $j = 1..J_k$
bus routes: $(R_l) = ((s^j)_l)$, for $l = 1..N_{routes}$ & $j = 1..J_l$
stops geographies: (s^j), for $j = 1..N_{stops}$
dissemination area geographies:
synthetic population: $(\pi)_j$, for $j = 1..N_{pop}$
trip chain model parameters

Result:
smart card data with socio demographic attributes

initialize the oriented object model;
set walking distance threshold: d_w;
load model parameters;
for $sm \in Smartcards$ **do**
  identify *home* as the most frequent departure stop;
  compute relevant trip chain statistics (average time of first departure, last departure, average number of daily activities, public transit usage etc);
  identify smart card's alternative group for the trip chain choice model;
end
for $st \in Stations$ **do**
  $localPopulation$;
  $localSmartcard = st$.getSmartcards();
  for $zn \in disseminationAreas$ **do**
    if $dist(st - zn) \leq d_w$ **then**
      $localPopulation$.add(zn.population);
  end
  for $\pi \in localPopulation$ **do**
    π.sampleChoiceSetFrom($localSmartcards$);
    π.applyModelOnChoiceSet();
  end
  constructCostMatrix($localPopulation$, $localSmartcards$);
  applyHungarianAlgorithm();
  for $\pi \in localPopulation$ **do**
    if π *has a smart card* **then**
      $population$.remove(π);
  end
end

Algorithm 3: Assigning smart cards to the synthetic population.

3.7 Validation

We want to validate the socio-demographic dimension of the outcome of our methodology (see 3.1). There are three different levels of validation possible: macroscopic validation, mesoscopic and microscopic. The best validation possible is at microscopic level, however to do so we need access to identified smart cards so we can check how our methodology performs in terms of socio-demographic characteristics imputation. To the best of our knowledge, the only work that had access to this kind of data is Munizaga et al. (2014). Spurr et al. (2015) proposed a method to identify smart card holder among the surveyed population based on a heuristic approach, however no validation of the methodology against ground truth has been done therefore we can't be sure about the outcome and we cannot use it to validate our work as it would be another layer of errors added. As we described in the section 3.5, smart card data is tied to a fare type which allows a minimum knowledge about smart card holder occupation (student, retiree, regular). This information can be used to make a partial validation of the methodology when the information is not primary used for the choice set generation (section 3.5).

Since ground truth is not available, microscopic validation is not a viable option. Therefore we decided to work with smart card data leveraged during the OD survey was held and to assume this would be the best proxy of the ground truth we could get for marginal distributions. There are two levels of validation on marginals: macroscopic and mesoscopic. First we can check that the global population using public transport is reproduced. In a second step, a more detailed validation can be done considering spatial information: is our method able to reproduce the spatial distribution of the STO user population?

The global consistency of the population affected to smart cards with the population reported to use the STO network is quantified using the Square Root Mean Squared Errors. It checks the internal consistency of the population with the reported population. We expect rather high values since we are computing the SRMSE over $age * gender * car * occ * nPers = 1120$ internal categories for a population of approximately 23 500 public transit users. Marginal distributions are checked as well.

We are using three different hypothesis to infer attributes: a) smart card users are living near their most frequent first bus stop of the day, b) they have a specific trip pattern related to their socio-demographic attributes and c) the population is well segmented by the fare system (we are assuming that 100% of students pay a student fare and the same assumption is valid for retiree and regular users). We can develop a sensitivity analysis of these hypothesis by comparing distributions of socio-demographic attributes for:

- (A) the OD survey
- (B) the smart card holder population while randomly assigned
- (C) the smart card holder population while using the local population assumption
- (D) the smart card holder population while using the local population assumption and the trip chain choice model
- (E) the smart card holder population while using the local population assumption, the trip chain choice model and the fare type

This approach helps understand whether hypothesis made perform better than a random affectation of smart card or not, and how much it improves.

The method we proposed has three independent steps and each of them can produce errors which then are transferred to following steps. Each of these steps should be monitored and validated so the global methodology has a chance to produce a good output. The most common validations for population synthesis are Total Absolute Errors (TAE), Standardized Absolute Errors (Anderson et al., 2014)(Ballas et al., 1999)(Huang and Williamson, 2001). Ballas et al. (1999) defines TAE as the absolute difference between estimated probability rates.

$$TAE^a_{x,i} = \frac{\tilde{n}^a_{x,i}}{\tilde{N}_x} - \frac{n^a_{x,i}}{N_x}$$

Where TAE^a_x is the TAE over attribute a, category i at location x, $\tilde{n}^a_{x,i}$ is the number of synthesized agent in dissemination area x which attribute a is the category i, $n^a_{x,i}$ is the number of person from the dataset in dissemination area x which attribute a is the category i, $\tilde{N}_x$ is the total number of synthetic agent in dissemination area x and N_x is the total number of person in dissemination area x.

Destination and activity location inference are difficult to validate because it also requires to be able to link smart cards to a person *a priori*. Usually they are not individually checked: they are validated at an aggregate level if the aggregate data match reasonable expectations (or analysis made out of the travel survey). For our work we are using smart card data and a methodology that was already used in previous work, therefore we recommend the reader to consult Trépanier et al. (2007) to know more about the validation process for stops alighting.

The trip chain choice model is calibrated on 80% of the OD survey and is then validated using 100% of the data available using confusion matrix and marginal checks. Model and parameters statistics are carefully analyzed.

CHAPTER 4 CASE STUDY: GATINEAU

The datasets were already introduced in the methodological section. Further description of the public transit network can be found in (iTRANS Consulting Inc., 2006) and (Blanchette, 2009). This section describe the three mains stages of our methodology: a) the population synthesis, b) the calibration of a trip chain choice model and c) the actual association between persons and smart cards.

4.1 Population synthesis

Population synthesis was operationalized using the open source software SimPSinz (Farooq, 2013)(Farooq et al., 2013c)(Anderson et al., 2014). The code is available in Java and C#. The Java version was used and adapted for our case study, especially on the optimization side: we implemented multi-threading to speed up computation and we implemented data batching to reduce RAM memory usage. One of our hypotheses is that there is a distance threshold between the closest bus station and the place of living over which the agents won't consider public transit as an available alternative. In our case study 83% of public transit riders live within a 1km radius from a bus stop. Therefore we synthesized a population only for dissemination areas that were within a 1 km radius of a station. It reduces the number of dissemination areas in the CMA of Ottawa-Gatineau (CMA code: 505) from 1781 to 552 dissemination areas. This avoids the population synthesis of unnecessary population and speeds up the population synthesis process. We also developed an analysis function which produces a goodness of fit analysis for each local area based on TAE, SAE.

We tried using Importance sampling to combine global distributions and local distributions whenever it was possible (age and gender), however local distributions can be widely different from global distribution and the random walk would never produce a sample which fit both distributions. Therefore we decided to use only Gibbs sampling and to substitute global distribution by local distributions whenever it is possible (age and gender). Since local distributions of age and gender are conditional marginal, it may result in producing some agents that are 'inconsistent' (for example a 10 years old kid working), but internal consistency is controlled by 6 distributions out of 8 and the gender distribution is basically a uniform distribution so it will not induce any inconsistencies.

For each batch of dissemination areas, and then for each computing thread, the Gibbs sampler was warmed up with a 500 000 draws. A 1000 draws would have been enough if we were doing

only one warm up for the whole process, however since we are working on multi-threads, we have to operate a warm up for each of them. Then we sampled an agent every 1000 draws (skip parameter). The method ran in 38 minutes, with 15 batches and 7 logical processors of a i7-4710 2.5GHz processor, and 8 Go of RAM. Note that the Gibbs sampling is an asymptotic method (Chan, 1993) therefore we could have use a smaller skip criterion.

At the local level, age and gender marginal distributions of the population synthesis fit well with marginal distributions from the census data. Figure 4.1 and 4.2 show SAE (standardized absolute errors) and TAE for age category 2 (25 to 34 years old) at dissemination area level. For most areas (around 500 persons), the total absolute error (TAE is beneath 15 persons which is quite low. When looking at SAE we see most of zones are beneath 0.2. Figure 4.3 and 4.4 show SAE and TAE for sex category female. On both criteria, gender attribute is better synthesized than age. It is mostly due to the fact that gender has only 2 categories versus 7 categories for age, and in addition the distribution among those 2 categories is almost uniform. For both age and gender, SRMSE is close to 0 for all dissemination area (see figures 4.5 and 4.6). Distributions of SRMSE (square root mean squared errors) (see figures 4.7 and 4.8) show that most of dissemination areas very well simulated since approximately 90% of dissemination areas have a SRMSE beneath 2 for gender and 1.2 for age which are very low values. It is interesting to note that dissemination areas where the population synthesis performed the worse are zones where the gender distribution is highly unbalanced.

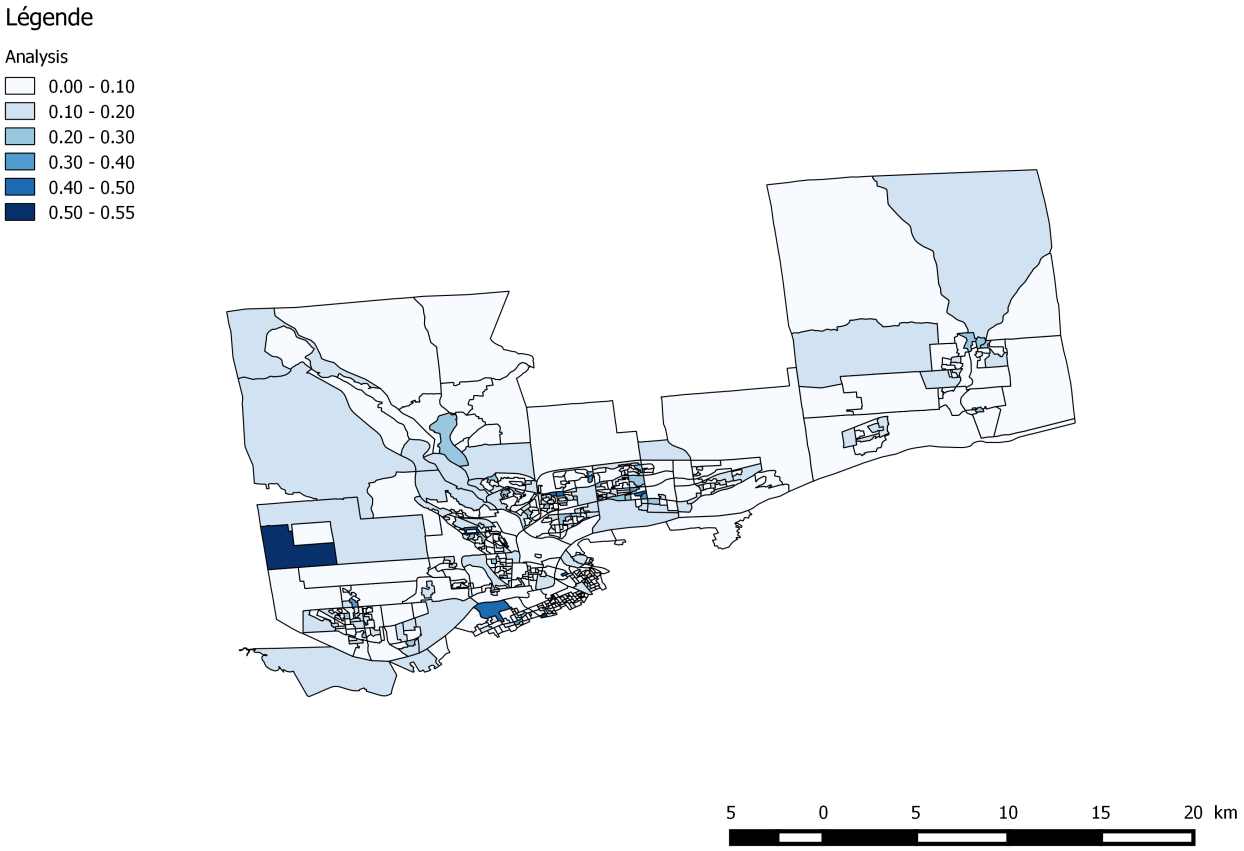

Figure 4.1 SAE for age category 2 (25 to 34 years old) at local level.

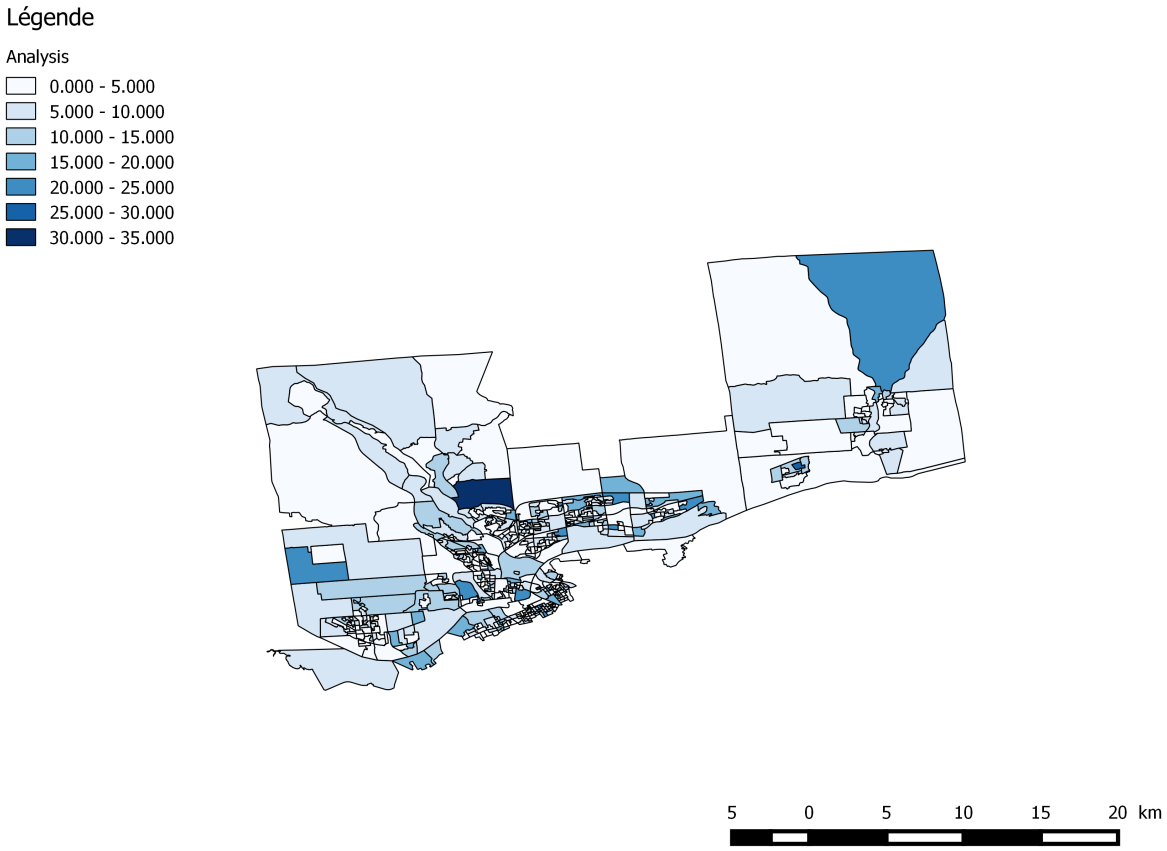

Figure 4.2 TAE for age category 2 (25 to 34 years old) at local level.

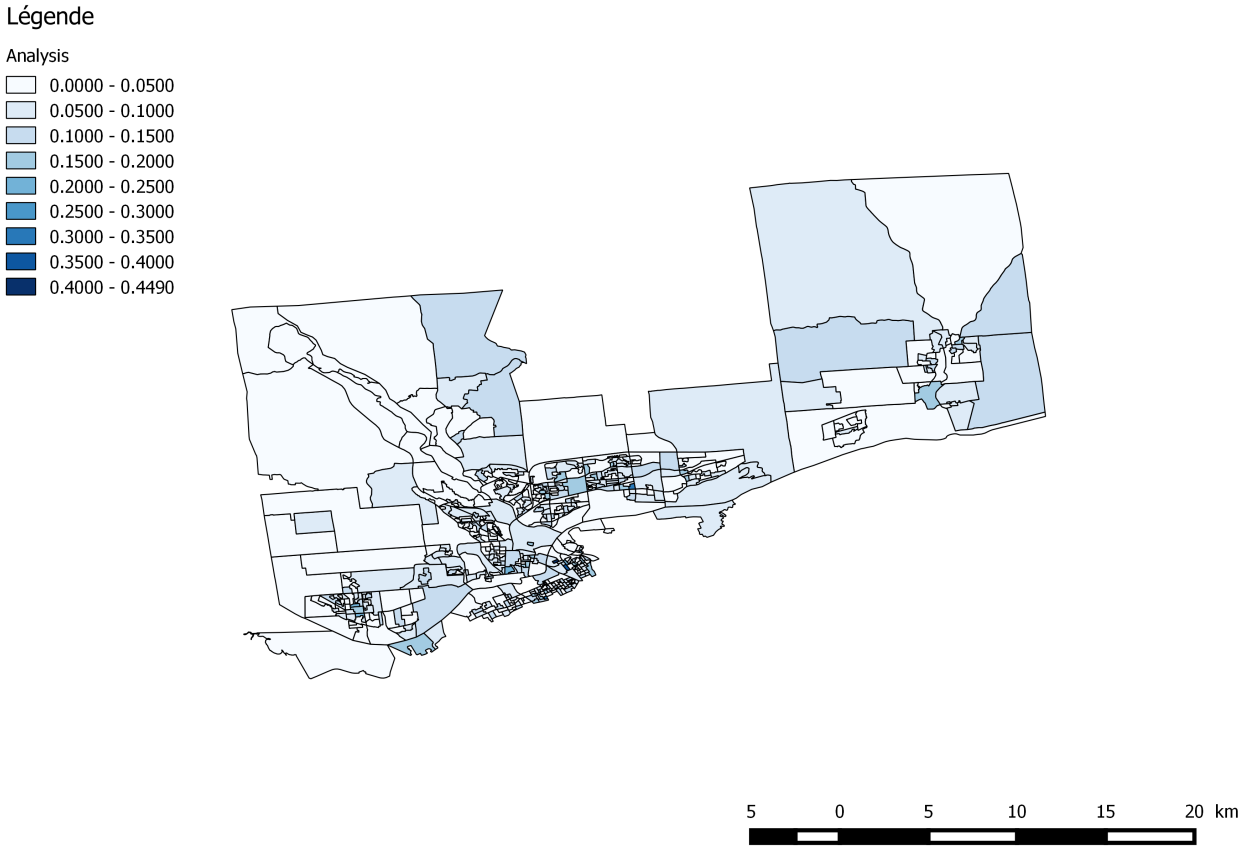

Figure 4.3 SAE for gender female at local level.

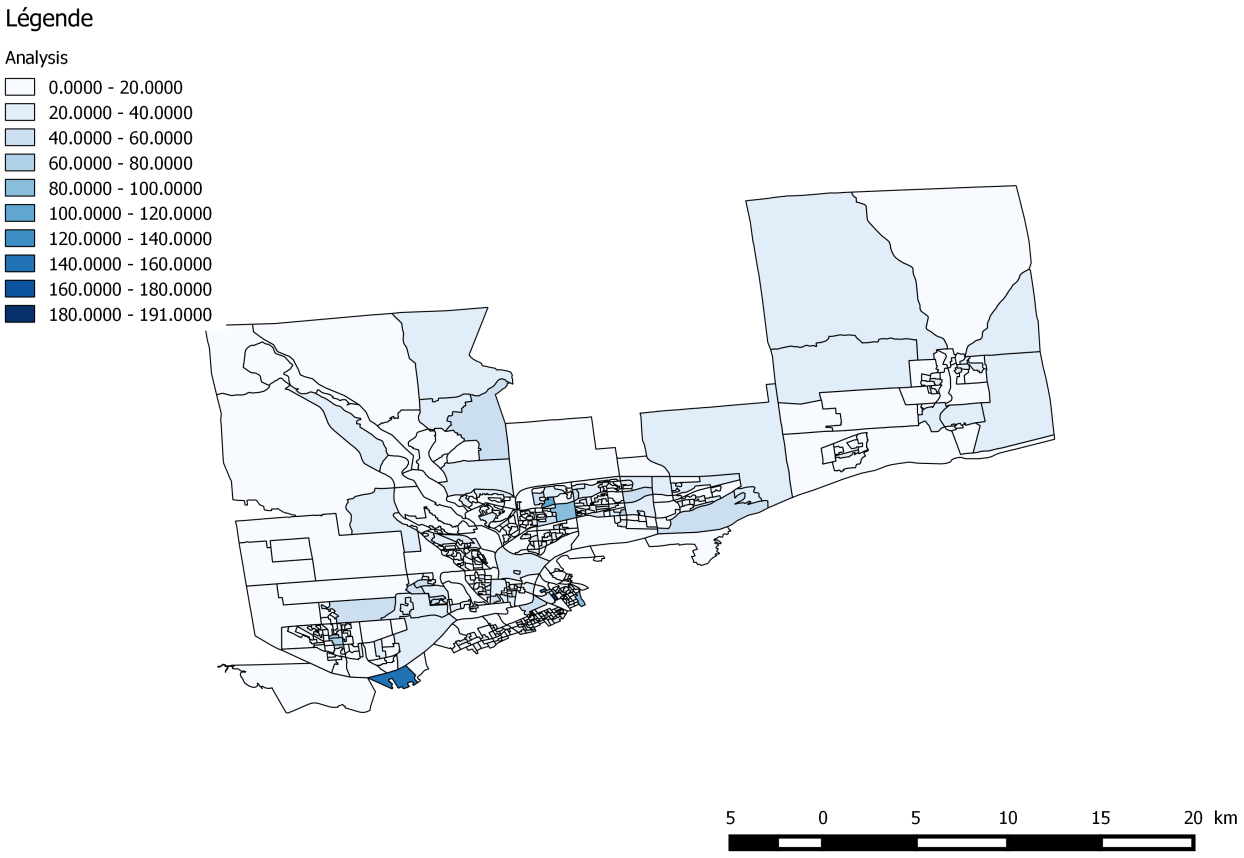

Figure 4.4 TAE for gender female at local level.

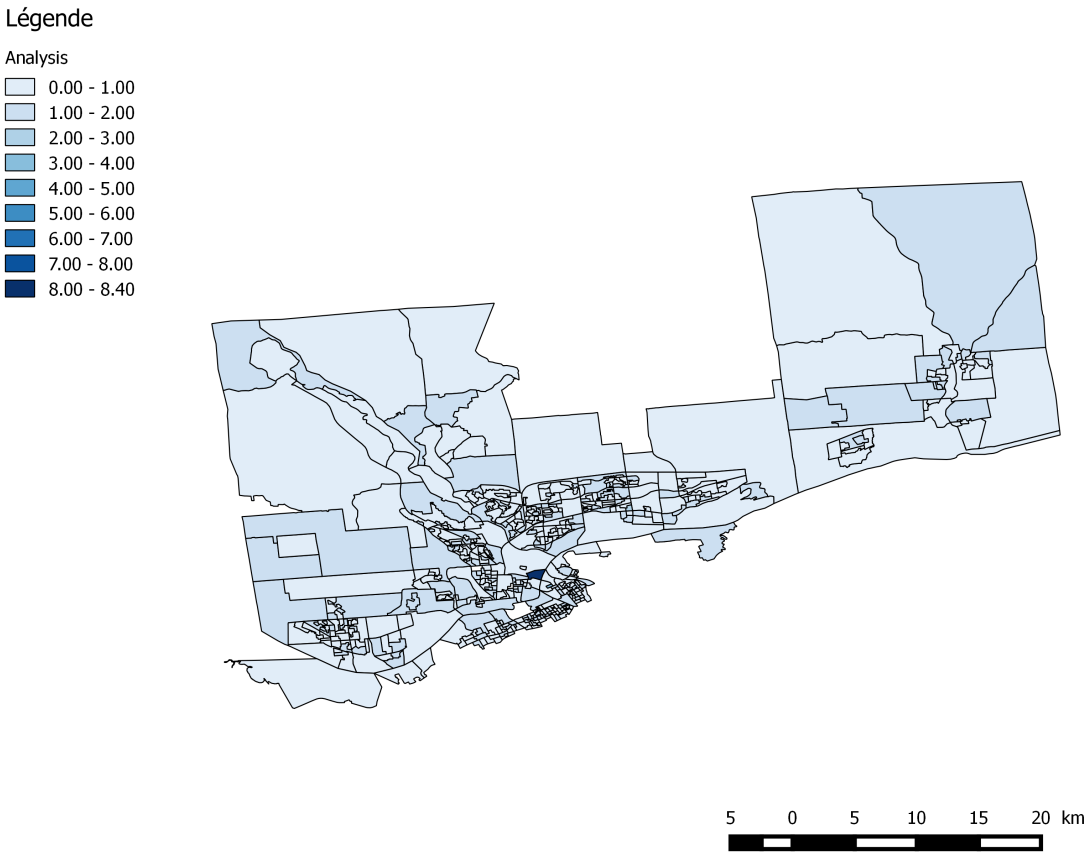

Figure 4.5 SRMSE for age distribution.

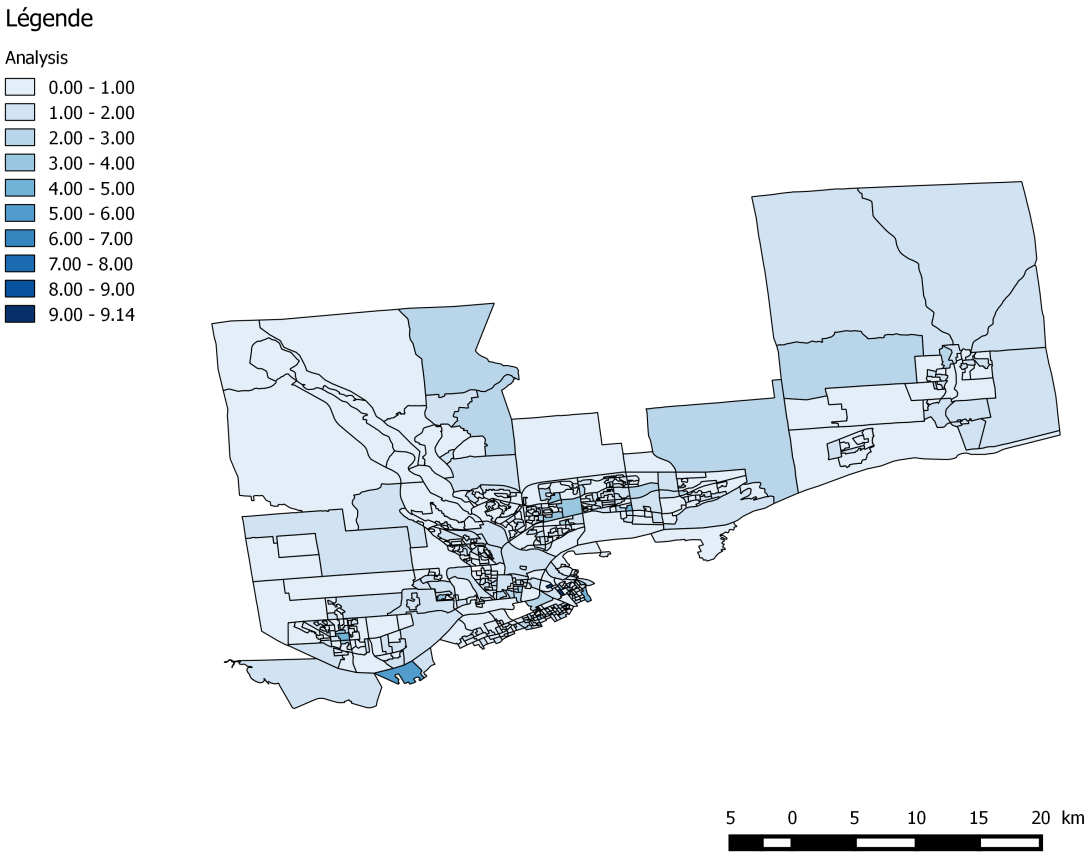

Figure 4.6 SRMSE for gender distribution.

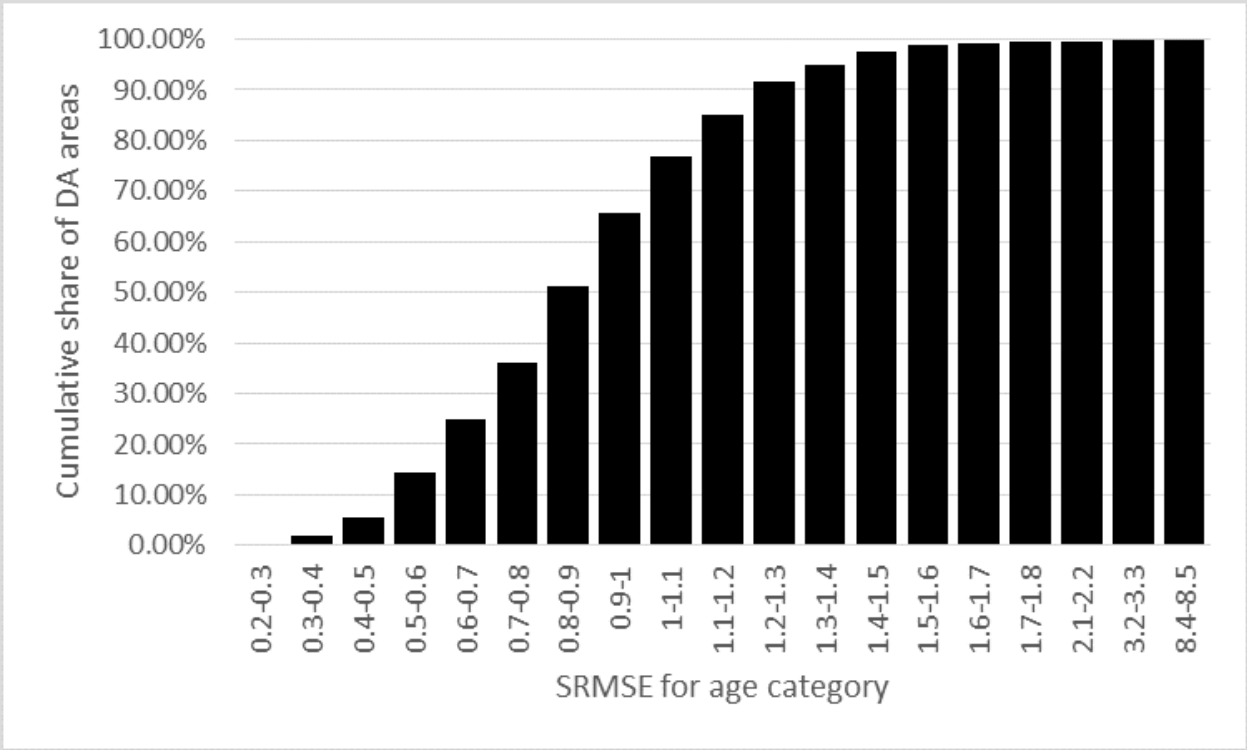

Figure 4.7 SRMSE cumulative distribution for age

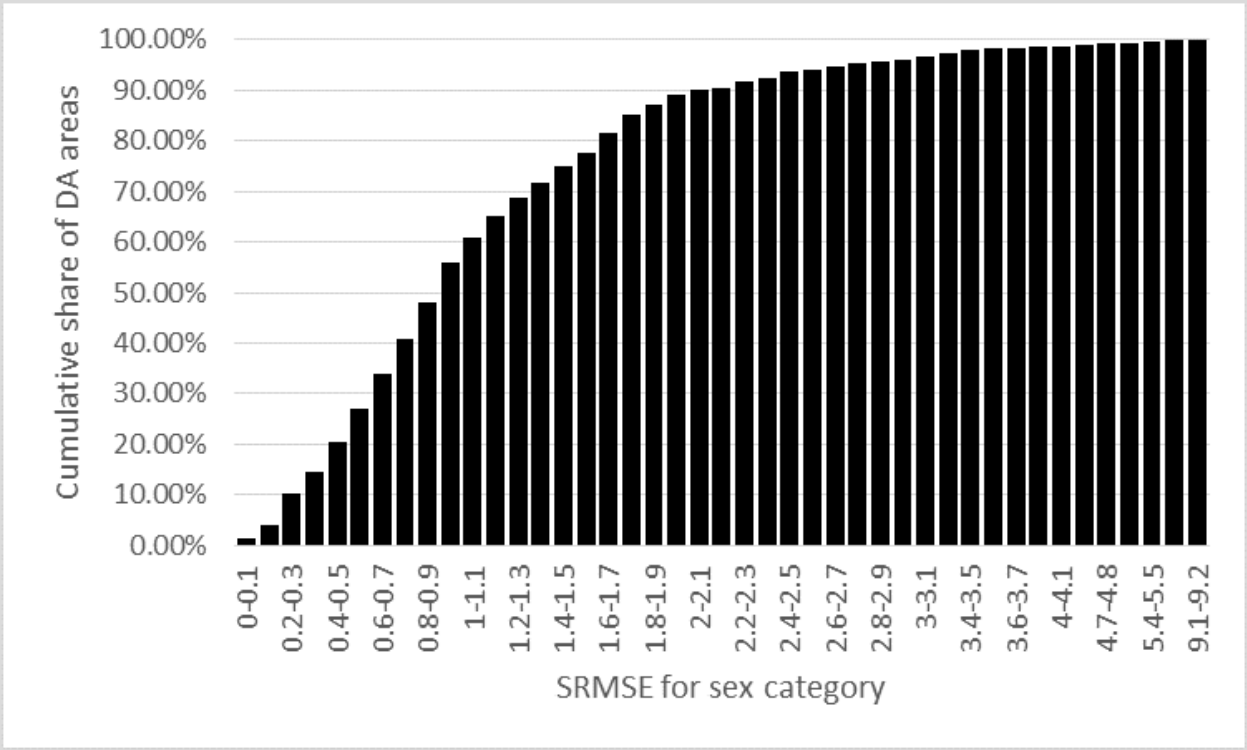

Figure 4.8 SRMSE cumulative distribution for gender.

4.2 Trip chain choice model

For each individual of the OD survey, we identified the trip chain choice according to the description made in the methodology part (base on time of departures, number of activities and public transit usage). Figure 4.10 shows that with our chain description, 10 types of trip chains contain approximately 90% of the mobility patterns observed through the travel survey. Those choices are mostly loyal public transit users (they didn't use any other mode to travel). Figure 4.9 describe each choice distribution over occupation. We can see that some categories are highly segregated: for instance more than 80% of people who chose C_2_1_1_0 (from left to right, number mean: person loyal to public transit - first departure hour during peak hour - 1 activity - going back home earlier than evening peak hour) are students.

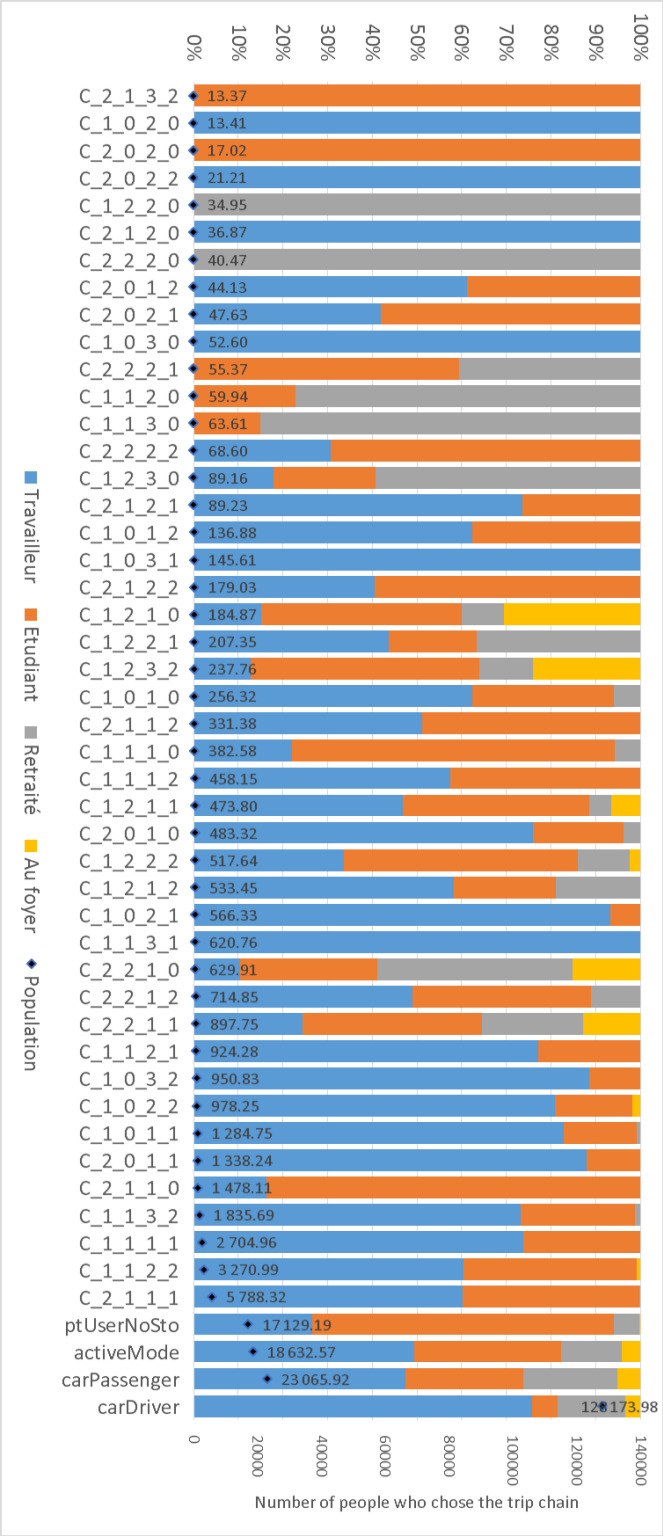

Figure 4.9 Occupation share for each choice, data: OD survey 2005. The format is C_PTusage_FirstDep_nAct_LastDep (more thorough description in Table 3.4. It is also sorted from the most frequent choice (up) to the least frequent choice (bottom).

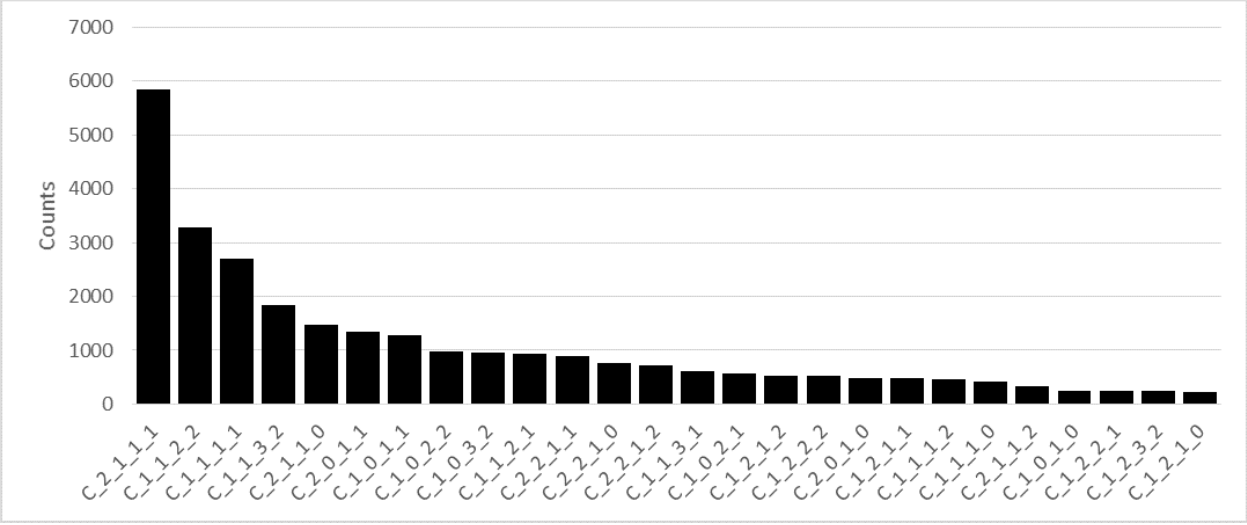

Figure 4.10 Distribution of the most frequent choices in the travel diary survey.

The trip chain choice model was estimated using the open source software Biogeme (Bierlaire, 2003) and control file for joint model estimation were automatically produced using a software we developed and have made available on a github repository (Grapperon, 2016). Given our framework, for the explanatory variables of the utility function, we can only use agent specific attributes and accessibility indicators at DA level since alternative specific variables are not known for unobserved trip chains. The following modeling assumptions were made (with respect to the descriptive analysis of exactly the same dataset available in Blanchette (2009) and by carefully looking at distributions drawn from the travel survey):

- most of public transit user who didn't use STO are kids using school transportation
- middle-aged person use mainly private mode since they have both the physical ability and the money to do so
- retiree don't use STO services since they are mostly used to drive car and when they are not able any more they are more likely to use some special service for people with disabilities
- women are more likely to use STO services or to be car passengers
- couples without children and single persons are more likely to use active modes since they have less time constraints
- elder generation don't use public transit since they are not accustomed to
- children use STO services
- retiree have fewer time and spatial constraints therefore they are more likely to travel out of peak hours.
- home-keeper are in the same case
- young people leave school earlier than peak hour
- the bigger a family, the more intra-household constraints there are, inducing a more constrained schedule, therefore people travel during peak hours

We used 80% of the travel survey data to calibrate the model and we simulated the model on 100% of the data. General statistics can be found in Table 4.1 and parameters statistics can be found in Table 4.2. General statistics gives fairly good results with a Rho-square stats of 0.663; which is pretty high for such a comple model. The nested structure is validated since

the scale factor for STO nest is higher than any parameter. Constants absolute values are quite high showing that there are still important phenomena remaining unexplained. The scale values were estimated on simpler models then we fixed them to help the calibration achieve significant results.

The parameters that we are using are significant (the p-value is under 0.01). The signs of the parameters are consistent with prior expectations. For example we expected that elder people (over 65 years) would not be likely to use public transit. The related parameter is -0.571. It was associated with the STO nest and the public transit non STO nest. The negative value is reducing the utility value of these nests. All parameter values are between -1.54 and 1.64 while constants are between -2.42 and 1. Constants and parameters have similar range of values, which means that our model is still missing important information (for example, land use information, accessibility measure, choice specific attributes such as trip length). The nests scales for non STO nests were found to be 1, while the scale for the STO nest was found to be 7.5. This is consistent with the fact that there are sub-choices only for the STO nest. There is a need to scale and zoom in only for the STO nest.

Table 4.1 General statistics for the trip chain choice model

Model	:	Nested Logit
Number of estimated parameters	:	71
Null log-likelihood	:	-50978.718
Init log-likelihood	:	-22639.480
Final log-likelihood	:	-17131.386
Likelihood ratio test	:	67694.664
Adjusted rho-square	:	0.663
Final gradient norm	:	+1.152e+000

Table 4.2 – Parameters and statistics of trip-chain choice model

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	BACK_FROM_SCHOOL	0.162	0.0204	7.95	0.00
2	BIG_FAMILY_1ST_DEP_PEAK	0.0636	0.0139	4.57	0.00
3	C_1_0_1_0	-1.38	0.0751	-18.42	0.00
4	C_1_0_1_1	-1.10	0.0679	-16.25	0.00
5	C_1_0_1_2	-1.41	0.0825	-17.10	0.00

Continued on next page

Continued Table 4.2: Parameters and statistics of trip-chain choice model

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
6	C_1_0_2_0	-1.74	0.149	-11.71	0.00
7	C_1_0_2_1	-1.23	0.0711	-17.36	0.00
8	C_1_0_2_2	-1.15	0.0686	-16.74	0.00
9	C_1_0_3_0	-1.60	0.102	-15.64	0.00
10	C_1_0_3_1	-1.38	0.0786	-17.56	0.00
11	C_1_0_3_2	-1.16	0.0688	-16.79	0.00
12	C_1_1_1_0	-1.36	0.0727	-18.72	0.00
13	C_1_1_1_1	-1.03	0.0672	-15.35	0.00
14	C_1_1_1_2	-1.27	0.0717	-17.66	0.00
15	C_1_1_2_0	-1.59	0.0930	-17.08	0.00
16	C_1_1_2_1	-1.17	0.0686	-17.01	0.00
17	C_1_1_2_2	-1.00	0.0670	-14.99	0.00
18	C_1_1_3_0	-1.63	0.100	-16.23	0.00
19	C_1_1_3_1	-1.23	0.0702	-17.52	0.00
20	C_1_1_3_2	-1.09	0.0680	-16.11	0.00
21	C_1_2_1_0	-1.54	0.0849	-18.10	0.00
22	C_1_2_1_1	-1.27	0.0707	-17.94	0.00
23	C_1_2_1_2	-1.26	0.0710	-17.73	0.00
24	C_1_2_2_0	-1.74	0.115	-15.13	0.00
25	C_1_2_2_1	-1.39	0.0765	-18.20	0.00
26	C_1_2_2_2	-1.26	0.0711	-17.66	0.00
27	C_1_2_3_0	-1.64	0.0951	-17.29	0.00
28	C_1_2_3_1	-2.39	0.0655	-36.51	0.00
29	C_1_2_3_2	-1.36	0.0758	-17.97	0.00
30	C_2_0_1_0	-1.27	0.0728	-17.45	0.00
31	C_2_0_1_1	-1.05	0.0672	-15.64	0.00
32	C_2_0_1_2	-1.50	0.101	-14.78	0.00
33	C_2_0_2_0	-1.71	0.150	-11.40	0.00
34	C_2_0_2_1	-1.55	0.115	-13.44	0.00
35	C_2_0_2_2	-1.64	0.149	-11.00	0.00
36	C_2_0_3_0	-2.39	0.0657	-36.36	0.00

Continued on next page

Continued Table 4.2: Parameters and statistics of trip-chain choice model

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
37	C_2_0_3_1	-2.31	0.0643	-35.94	0.00
38	C_2_0_3_2	-2.31	0.0643	-35.94	0.00
39	C_2_1_1_0	-1.13	0.0671	-16.86	0.00
40	C_2_1_1_1	-0.882	0.0659	-13.38	0.00
41	C_2_1_1_2	-1.28	0.0732	-17.51	0.00
42	C_2_1_2_0	-1.65	0.118	-14.00	0.00
43	C_2_1_2_1	-1.43	0.0859	-16.63	0.00
44	C_2_1_2_2	-1.35	0.0769	-17.53	0.00
45	C_2_1_3_0	-2.42	0.0662	-36.53	0.00
46	C_2_1_3_1	-2.34	0.0646	-36.18	0.00
47	C_2_1_3_2	-1.67	0.148	-11.23	0.00
48	C_2_2_1_0	-1.35	0.0750	-17.99	0.00
49	C_2_2_1_1	-1.15	0.0681	-16.91	0.00
50	C_2_2_1_2	-1.19	0.0706	-16.83	0.00
51	C_2_2_2_0	-1.72	0.113	-15.22	0.00
52	C_2_2_2_1	-1.58	0.114	-13.89	0.00
53	C_2_2_2_2	-1.53	0.102	-15.03	0.00
54	C_2_2_3_0	-2.58	0.0699	-36.92	0.00
55	C_2_2_3_1	-2.34	0.0649	-36.11	0.00
56	C_2_2_3_2	-2.34	0.0649	-36.11	0.00
57	C_activeMode	-1.35	0.0629	-21.53	0.00
58	C_carPassenger	-1.36	0.0622	-21.87	0.00
59	C_ptUserNoSto	-1.99	0.0776	-25.69	0.00
60	ELDER_DONT_USE_PT	-0.571	0.0954	-5.98	0.00
61	HOMEKPR_1ST_DEP_LATE	0.370	0.0417	8.89	0.00
62	MID_AGE_USE_PRIV_MODE	0.499	0.0547	9.12	0.00
63	RET_DONT_USE_PT	-1.54	0.153	-10.07	0.00
64	RET_GO_HOME_EARLY	0.307	0.0437	7.03	0.00
65	RET_LEAVE_HOME_LATE	0.457	0.0456	10.00	0.00
66	SCHOOL_TRANSPORTATION	1.64	0.0891	18.40	0.00
67	SMALL_FAM_USE_ACTIVE	0.132	0.0705	1.87	0.06

Continued on next page

Continued Table 4.2: Parameters and statistics of trip-chain choice model

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
68	WOMEN_ARE_PASS	0.853	0.0638	13.36	0.00
69	WOMEN_USE_PT	0.479	0.0561	8.55	0.00
70	WORKER_ARE_NOT_FIDEL	0.0690	0.0134	5.15	0.00
71	YOUNG_USE_STO	-0.206	0.0739	-2.78	0.01
72	stoUser	7.50	.	.	.
73	carDriver	1.00	.	.	.
74	carPassenger	1.00	.	.	.
75	ptUserNoSto	1.00	.	.	.
76	activeMode	1.00	.	.	.

The model was used on 100% of the of the travel survey population. Marginal check for each choice was performed and the confusion matrix was drawn. It is a 58 x 58 table therefore it won't be displayed here. First versions of the model were not performing well to reproduce nests population and we had to come up with a few constraints to improve the results. At first the use of car driving was not well reproduced for the age category 11 years old to 19 years old because it is a category including people that are not allowed to drive and people that are allowed to drive. To compensate this issue due to a bad age category definition, we set that one out of twenty individuals in this category would be given the choice of car driving. The rational is that only people 16 years old with a driving license are allowed to drive. We also forbade people not owning a car to use car driving mode. It represents a very small share. The age category 11 to 19 years old is using STO service or public transit service (school transportation) a lot, but the model is not able to make a difference between those two nests. It resulted in over estimating public transit other than STO and under estimating STO share. That is why, for age category 0 (11 to 19 years old), whenever the STO nest or the public transit no STO nest is chosen, we redistribute the choice randomly between those two according to the observed distribution.

The final model is performing quite well (see marginal distributions in figure 4.11 and 4.12), however it is over estimating STO mode for age categories 5 and 6 (over 55 years old) and under estimating car passenger in a similar manner. This is due to the fact that we are lacking of data and explanatory variables to be able to calibrate a better model. There are important phenomena remaining unexplained. The biggest nest (car driving) is well reproduced and for

smaller nests, the global shape is respected while marginal counts are close to the observed counts. There are exceptions: the model does not perform well to differentiate public transit no STO nest from car passenger and active mode. They are smallest nests and they can be seen as close alternatives as they all can be seen as alternatives to the car driving mode. The confusion matrix shows that between 40% and 50% of simulated choices are identical to observed choices (mostly thanks to the car driving mode). In our model we focused on agent specific attribute while we could have used more chocie specific attributes. For example, travel distance could be a good attribute to include to model active modes.

We do not present an analysis of the model for each case of the STO users nests since it would mean to explain 54 cases. The marginal analysis we have presented is considered here as a sufficient validation to carry on with our work.

Table 4.3 Confusion matrix of the trip chain choice model

	sto users	active	car driver	car pass	pt users
sto users	**0.147**	0.125	0.44	0.13	0.14
active	0.16	**0.13**	0.37	0.16	0.18
car driver	0.08	0.07	**0.69**	0.07	0.09
car pass	0.12	0.10	0.54	**0.11**	0.13
pt users	0.23	0.20	0.18	0.17	**0.21**

Figure 4.11 Marginal distribution of age, gender, occupation, household size and number of cars observed in the OD survey.

Figure 4.12 Marginal distribution of age, gender, occupation, household size and number of cars simulated for OD survey agents.

4.3 Association

We used the Hungarian algorithm implemented in Java from the SimPSynz software (Anderson et al., 2014) to be able to match the population to smart card data. For technical reasons (lack of RAM memory) we downsized the walking distance threshold from $1km$ to $500m$. Egu (2015) made a sensitivity analysis of this criterion and found it was not making a significant difference. We expect that the walking distance threshold is a value that could be finely tuned with an accessibility model to the public transit system. However, in our case study, we are reducing the walking distance threshold out of computational requirements. In general, our software is able to come up with a solution for local populations under 10 000 persons (in the Gatineau case with a $500m$ walking distance threshold, figure 4.13 shows local population size). Multi-threading is impossible since the population pool would have to be a dynamically shared variable, and in addition multi-threading would multiply RAM memory requirement by the number of threads. Another way of downsizing RAM requirements would have been to improve the Hungarian algorithm to enable him to not consider void cases. The association process runs in approximately 8 hours for 28 000 smart cards and a population of 300 000 persons. Local populations around a station are in average 4 000 and the average number of smart cards associated to a station is 11, close to half the stations have no smart cards associated with them. The major explanation is that the STO network in 2005 is mainly used for commuting from suburbs to major trip attractors. Therefore at a specific location, one bus stop will be associated with many smart cards while the bus stop on the other side of the road is the place where people get off at the end of the day (see figure 4.14). Among stations with smart cards, the average local population size is 4 300 and the average number of smart cards is 20. Local population sizes range from 2 000 to 8 000 and local smart card counts range from 0 up to 650. Three stations have more than 500 smart cards attached. Two of them are hubs with a car parking incentive (it represents 1 200 smart cards) and it is likely that a significant part of those smart card holders did not walk to the station but rather drove there or were chauffeured. This is a limit of our first hypothesis and an incentive to come up with an accessibility modeling approach more refined than a simple distance threshold.

In order to know more about the sensitivity of our methodology, we ran four different types of associations:

- random distribution of smart cards to the global population
- random distribution of smart cards to local population (within the 500 m buffer) around each bus stops

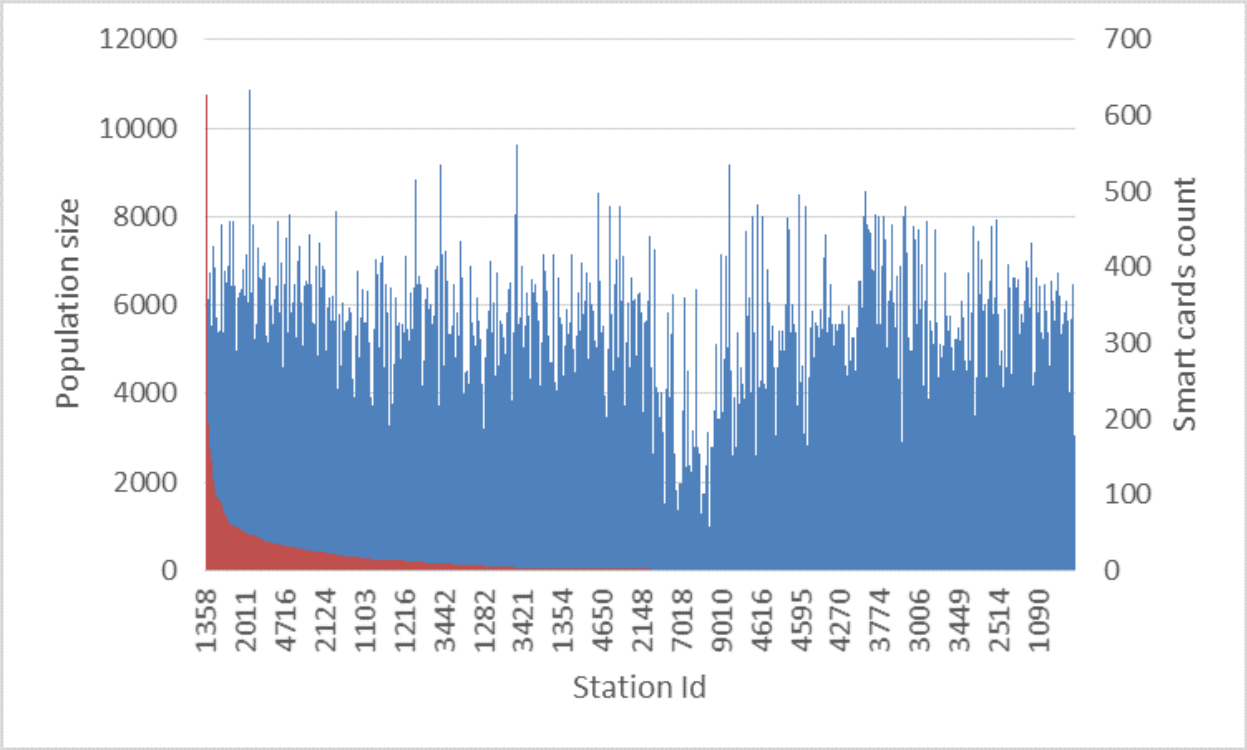

Figure 4.13 Local population sizes (blue) and count of local smart cards (red) around bus stations.

- distribution of smart cards to local populations using the trip chain choice model approach
- distribution using local population, trip chain choice model and the fare type criterion..

4.3.1 Macroscopic validation

Resulting STO users population are drawn in figure 4.15. We also drew the observed population from the OD survey (only 80% since that is the penetration rate of the smart card amongst the STO users). We did a random distribution to the global population to have a base reference: is our methodology performing better than no methodology at all? When looking at the distributions, the reader has to keep in mind that the 20% of the OD survey are not represented in the smart card data. They may be a mixed population of non smart card users and smart card users from the OC transport network (Ottawa's public transportation network). They may be driven by a specific pattern (for example, maybe old people don't use smart card) and therefore assuming smart card NON-ownership is homogeneously distributed among the population would induce biases. The STO user population is rather well reproduced. However it shows multiples tendencies: it is over estimating the number

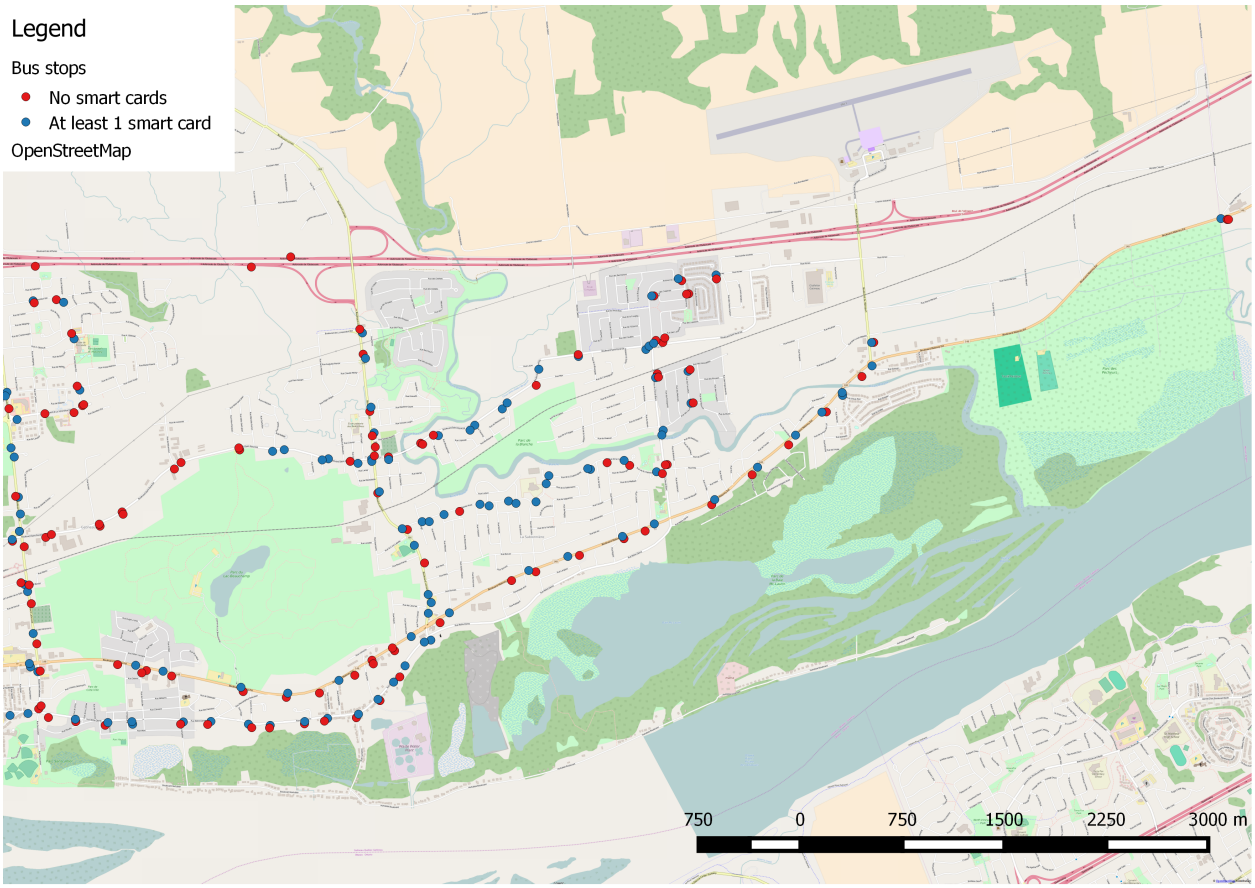

Figure 4.14 Bus stops along boulevard Malorney.

of single person who uses STO. It is not performing well for number of cars since the global random distribution works better than any other solution, here there is a phenomenon that was missed in our model. By comparing student share from the OD survey and from the smart card data (who paid a student fare), we can see that there is really a higher share of students than what is observed from the OD survey. This support the fact that we need a smart card ownership model. The number of student owning a smart card in the OD survey should be close to the number of student smart card found in the smart card database. But we have no ways to check other occupation types and to create a smart card ownership model. The unbalanced gender distribution in the STO user population is not well reproduced since our results show that there is close to no difference whereas the OD survey has a significant difference. The age distribution is rather well reproduced, except for the age category 1 (20 to 24 year old) and for the age category 6 (other 65 years old).

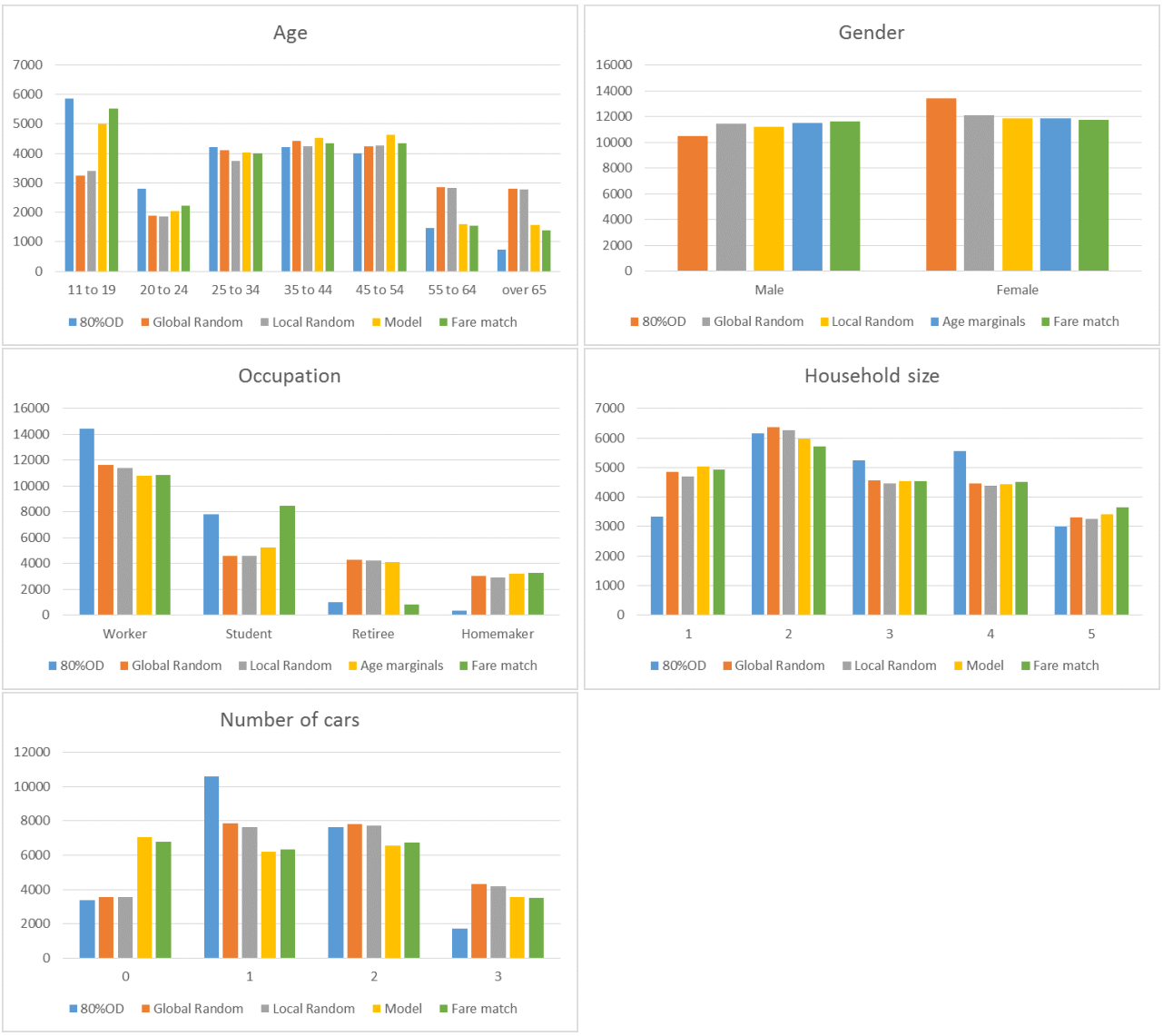

Figure 4.15 Marginal distributions of age, gender, occupation, household size and number of cars for distributed smart cards.

4.3.2 Mesoscopic validation

Another important marginal check can be done at a more local level. We select a geographical area which is segregated from the rest of the region by natural constraints and highways (see figure 4.16). In this area we expect our model to be able to recreate the local distributions of STO users. If it does, this would validate our methodology with a high degree of confidence. The first note is that our methodology assumes that there are 236 smart card users in this neighborhood while the travel survey states that there are 500 STO users. With the peneration correction, it makes 400 smart cards. The smart card counts don't match. It

could mean that the OD survey is not accurate enough. However it has to be understood that this is broad approximation since penetration rate of the smart card is not something equally distributed. There is an important difference of smart cards in the neighborhood compared to the travel survey. This can be explained by the very blind hypothesis we are doing to estimate that the most frequent first boarding of the day is the closest to the house. A deeper analysis of the travel survey data shows that approximately half of the population in the neighborhood who use STO network is also presenting a complex trip chain pattern (they are using other means of transportation). In a residential area of Gatineau, it is highly likely that they are commuting in the morning using car pooling and going back using STO in the evening. From figure 4.17, we can see that our methodology affects the age attribute following a distribution that is shaped in a similar way than what was observed from the travel survey. The missing smart card are mostly related to the youngest age categories (11 to 24 years old) which is the category most likely to be chauffeured by their parents. Distribution of household size and number of cars are very poorly reproduced. The gender distribution does not reflect the fact that women use more public transit than men. The distribution of occupation is rather good, however it is estimating that homemaker (those main occupation is taking care of the familial home) are using STO in this neighborhood while it is not the case. To summarize, age and occupation are attributes that are rather well reproduced which is consistent with the fact that we developed our activity choice model with more care for those two categories. Other attributes distributions are not performing as well because a) the trip chain choice model is lacking of explanatory power and b) the Hungarian algorithm is a deterministic distribution algorithm which may systematically favor a specific socio-demographic profile (therefore it reduces the heterogeneity of the human choice) and c) the population synthesis was spatially controlled over age and gender only, therefore the distribution of the other attributes may be rather far away from the ground truth and from the OD survey reported truth.

Table 4.4 Square Root Mean Standard Error results for the joint distribution of age * gender * occupation * number of cars * household size.

Methodology	SRMSE
Global Random affectation	2.410
Local Random	2.423
Trip chain choice model	2.409
Fare matching	2.280

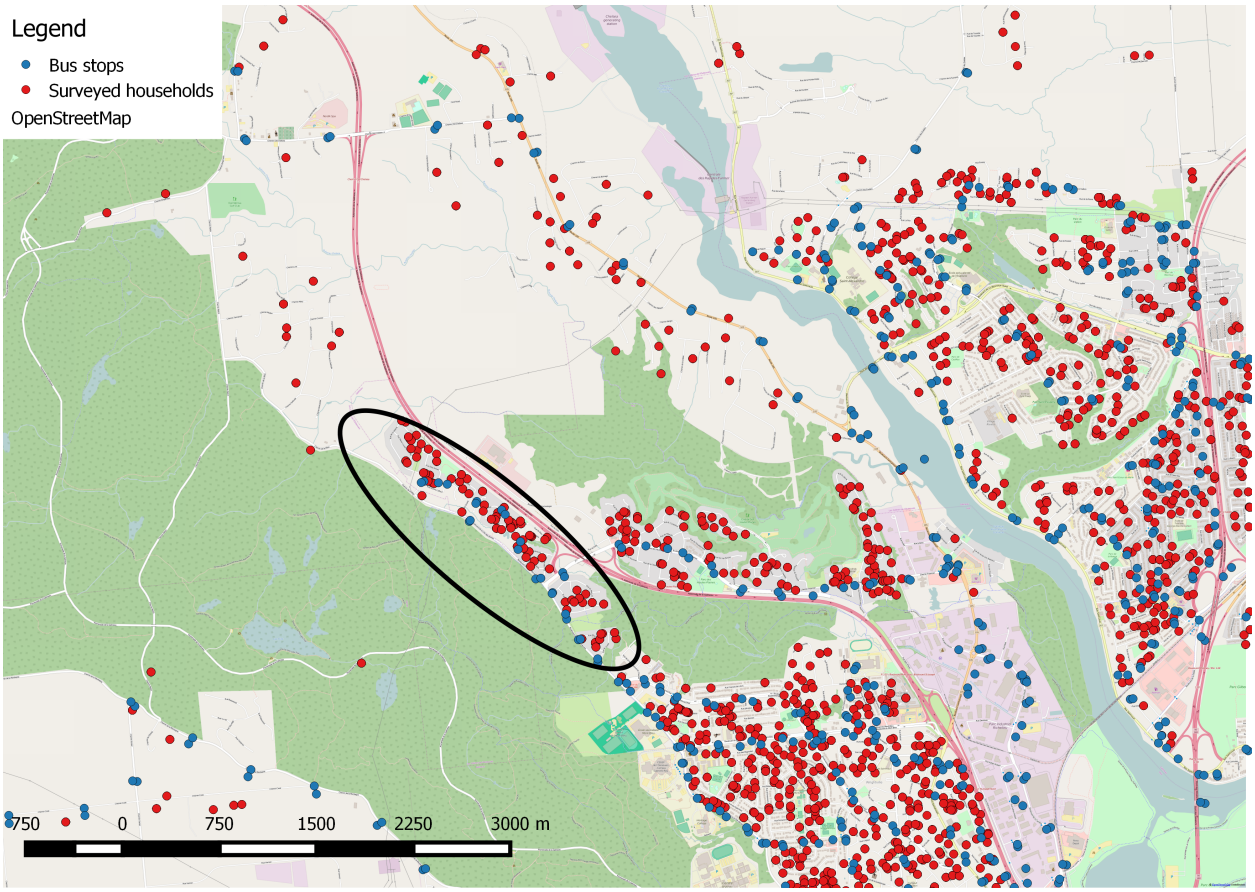

Figure 4.16 Geographical area selected for mesoscopic analysis of the results

4.3.3 Validation of the internal consistency of the socio-demographic dimension

The SRMSE indicator provides an estimate on cross-sectional counts and therefore it is an indicator that represents whether the internal consistency of the distribution within the socio-demographic dimension. We can see from the results (see table 4.4) that there is a slight improvement especially when using the third hypothesis (fare matching). But the activity choice model and the location hypothesis do not seem to improve the SRMSE measure. This reveals that it is really important to have at least one socio-demographic attribute to significantly improve the internal consistency of the results.

Figure 4.17 Local marginal distributions of age, gender, occupation, household size and number of cars for distributed smart cards.

CHAPTER 5 CONCLUSION

This section summarizes our work, then we highlight limits, and eventually we describe possible future work to overcome those limits.

5.1 Summary

We proposed a methodology to attach socio-demographic characteristics to smart cards based on three simple hypothesis: public transit users live close to the public transit network (within walking distance), they have different travel behaviors depending on their socio-demographic characteristics and they are understanding perfectly the faring policy. The methodology has three stages: a) synthesizing a population with a high spatial accuracy, b) estimating and calibrating a nested-joint trip chain model and c) weighting links between the synthetic population and smart cards and solve the resulting association problem. We created a synthetic population using a Gibbs Sampling method and validated it at the dissemination area level with traditional validation techniques such as Standardized Absolute Error, Total Absolute Error and SRMSE. We developed a nested-joint model for trip chain choice and calibrated it with Biogeme. We defined carefully the choice set generation to insure no kid could drive a car and we proceeded to some post calibration adjustment because the model was too weak to differentiate small nests from each other (the car driver nest was the only one which was clearly an independent choice from the other nests). Then we proceeded to the actual smart card enrichment with socio-demographic attribute by matching the synthetic population to smart cards using the graph theory Hungarian algorithm and the trip chain choice model to weight the links.

We ran a sensitivity analysis, comparing the output of a random association of socio-demographic characteristic from the global population, from the local population, from the local population using the association method based on trip chain choice model and from the local population using both the trip chain choice model and the applied fare. For each stage of our work a validation process was applied. Lacking the ground truth, as it is the case in many study based on smart cards, we assumed that the OD survey as a proxy since it was realized in the same time window than the smart card data. We also run a sensitivity analysis of our methodology to the various hypothesis. Limits are identified, some are related to the methodology, other are related to the data we used and others are related to hypotheses.

5.2 Limitation

There are many limitation which rose during the research. We pointed them out, and we develop here a critical analysis of our work.

5.2.1 Computational requirement

Most of the work presented was implemented through programming in Java. The population synthesis is relying on a Monte-Carlo Markov Chain random walk which can be exhaustive in computation time since the randomness of the walk has to be insured through a high repeatability of random steps. This time consumption was reduced through multi-threading. The Gibbs sampler runs on 7 cores with fairly good specifications in approximately one hour. Including Importance sampling could make the random walk wander *ad vitam eternam* if the distributions used are too different from each other. As it is coded, the association part is highly demanding in RAM because the Hungarian matrix requires a cost matrix linking all elements from one set to the other set, therefore a lot of links are null. It could be improved with more thoughtful code. The probabilistic association algorithm is more parallelizable and it also could help

5.2.2 Smart card data

The data quality is highly important. We used a prepared data set therefore we are assuming that the data quality is good and that errors are limited. Among possible errors: the smart card transaction can be attached to a wrong vehicle or to a wrong route, the clock attached to the faring device may be delayed and attach a wrong time stamp to the transaction. This kind of errors in the data can be hard to detect and to correct, that is why the goodness of the faring system is important so we can rely on the data it produces.

The rule based approach we used to detect alighting stops and infer activity location has the downfall of being rigid. It can't capture all the variety of individuals. This heuristic approach relies on very valid ground for the alighting stop inference: the physical capacity as a distance threshold. However we feel like the heuristic approach for inferring activity location is less legitimate. It is an interesting way of finding main activities since most of them are work or study and they usually happen at a specific location. However this approach based on a time threshold can't fit well with the variety of small activities a person can engage in. For this reason we think our definition of trip chain in our trip chain choice model should be less descriptive and more hardware driven using statistics that rely less on assumptions (such as the number of activities) and more on what is directly available (such as the number

of boardings). In our case it means that we could have used number of boardings instead of number of activities to described the trip chain because it is directly available from the dataset while number of activities is inferred from a rule based approach and this induces errors.

Another limit to using smart card data is the penetration rate. Gatineau offers a high penetration rate of approximately 80%, but it remains that the behaviour of the 20% is unknown and it can be different from the smart card owners. Therefore, any statistics based on smart card data should be handled carefully.

5.2.3 Population synthesis limitation

For the population synthesis, we used three different datasets which required to be very careful so the distributions are consistent with respect to spatial boundaries definitions and attributes definitions. For the nest of STO users, the travel survey was weak in providing information because the modal share is very low. The information spatially distributed is even scarcer. Socio-demographic information was both therefore we completed those two aspects with the census data and the PUMS. Importance sampling was not used to enrich local distributions with global distribution because the differences of distributions between the two scales are two high. A way to improve our population synthesis would have been to use Importance sampling for global distributions: the distribution of household size is known through the OD travel survey and through the PUMS data set. Another way would have been to develop discrete choice models for each missing attributes (Farooq et al., 2013c), which is also a good way to combine data with various levels of aggregation.

The population synthesis was validated at local level using classic marginal checks (Total Absolute Errors and Standardized Absolute Errors). The internal consistency of the produced agents is ensured by the methodology.

5.2.4 Trip chain choice limitation

Only 5% of the population used public transit in Gatineau in 2005, but a significant portion of those travelers are children using school transport and they are not STO users. Eventually our sample of STO users is quite small (1786 individuals) to estimate utility functions for our 58 choices.

The nested joint model was estimated using Biogeme. Model statistics are good and the nest structure is both motivated by the choices description and the nest scale value. It is a rather complex model with a nested-joint structure. Unlike most of actual trip chain model, we

put the mode choice in the upper layer. This helped us overcome computational limitations however it is probably performing less efficiently than it we had reversed the two layers of nests.

The choices were described using four alternative specific attributes. We decided to use only agent specific variables and accessibility indicators in our utility functions since there are already four alternative specific attributes involved. We could have added more alternative specific attributes which could be highly meaningful (number of transfers, trip chain length, indicators of local level of service etc) but since the final goal is to attach socio-demographic to smart cards, we decided to focus on that side. Important socio-demographic variables were not used in the trip chain choice model (income, education and marital status) since the information is not available in the OD survey. Usually more recent surveys have more explanatory variables available. However the lack of information could be overcome using a model to attach lacking socio-demographic attributes to the OD survey. It has to be noted that not all hopes should be founded on more explanatory variables since it can be expected that most of them are correlated and won't increase the model value.

Also, a limitation is that we do not consider activity location in our model. This could be done using land use data. Activity location is implicitly considered since we are using observed trip chain (therefore it is spatially defined). However it would be better for model results it we could consider activity location characteristics.

In the STO nest, we are modeling a nested-joint model to benefit from both structures. In the STO nest we are implicitly assuming that choice alternatives are Independently and Identically Distributed which we are not sure about. As we described in the literature review: most models imply that there is a given causal structure and operate nests between choices (for instance: first choice is a daily number of activities then the second choice is the time of departure) but these kind of model do imply a causal structure that may or may not be there, depending on each individual. A second consideration in favor of the nested - joint model is that we are given complete trip chain where all choices are already made, therefore it is much more convenient to implement a joint model whose output is a trip chain.

The trip chain choice model was calibrated on a travel survey which represents an average week-day. Therefore we cannot use fully the longitudinality of the smart card data (for instance week days and week ends choices, variation in the behaviour etc). This could be overcome using a longitudinal data set from another similar city: making the assumption that general behaviours are quite similar with respect to city structure and socio-demographic profile. However, spatial transferability is a complex process in itself and comes with its own issues. The travel survey we used is based on a proxy respondent: a single person in

the household provides the information for the whole family. Proxy respondent surveys are known to be highly biased.

5.2.5 Association limitation

Due to hardware requirement, the association part was not implemented as it was theoretically described. We had to do it sequentially in three batches of bus stops (it resulted in three batches which population was approximately 7000). It is still a complete distribution since we are removing assigned agents and smart cards after they were assigned, however the order in which they are assigned is not controlled. The RAM memory is the main limit to implementing the Hungarian algorithm over the whole population. We overcome this problem by batching the smart cards by geographical area. It makes sense since when two bus stops are far away from each other, there is no overlap between the local population of each bus stop. Another way would be to adapt the Hungarian algorithm to handle efficiently 'empty cases' (ie: when a person is too fare away from a bus stop, it is useless to compute and store a cost value for corresponding link).

The Hungarian algorithm is a very simple association methods. Other methods exists and they are able to consider the fact that we are using a probabilistic model to weight our links. Farooq et al. (2013d) describe an implementation of a probabilistic association method instead of a deterministic association method (as the Hungarian algorithm). The same kind of association method is more parallelizable. It would also reduce the RAM required and therefore we could implement our methodology in one shot instead of a) simulate a STO user population and b) distribute the smart cards amongst the STO users.

Besides the computation requirements, there are important limits to the assumptions we made. We assumed people have a single place of living which may be false (for instance for children with separated parents). We used a walking distance threshold to infer local population around a bus stop, however in some cases people drive or are chauffeured to the bus station, in those cases the access distance is greater. We can expect a greater amount of those park'n ride and kiss'n rides around larger bus station with a car park incentive. An improvement to our methodology would be to consider an access distance threshold instead of a walking distance threshold. The access distance threshold should be defined for each station as a measure of accessibility (walking or driving).

We defined the place of living of individual as the bus stop where there is their most frequent first transaction of the day. However for various reasons people can use a different mode to go to their place of activity (work, study...) and go back home using public transit. Therefore our methodology can infer the wrong place of living.

5.2.6 Summary of sources of errors

Our work is combining various theoretical and practical aspects and each of them brings a risk of errors. Table 5.1 gives a summary of those sources. The OD survey in itself is the source of multiple errors that spread over the population synthesis, the model calibration and the validation process. The 2005 Gatineau OD survey has very few explanatory variables and the public transportation service is diversified (there is the STO service, school services, Ottawa transportation authorities and many others) therefore the information about STO usage is only partially reliable. The nested-joint descret choice structure assume that mode choice is made before trip chain choice. This structure is fitting perfectly with our problem, however it is going against traditional nested structure (choice of activities location and time, then mode choice). Smart card data is not completely reliable due to hardware failures or misuses of the information system by bus riders and users.

Table 5.1 Sources of errors

Source	Error type	Impact
OD survey	proxy respondent	strong
	lack of relevant information	strong
	sampling strategy	weak
PUMS	proxy respondent	weak
	sampling strategy	weak
Population synthesis	internal consistency	weak
	spatial distribution based on global information	medium
Trip chain choice model	goodness of fit	strong
	too simple nest structure	strong
	mode choice before schedule choice	medium
	not activity location modeling	strong
	not considering land use	strong
	weakness of accessibility indicators	strong
Smart card data	AFC/AVL system failure	weak
	smart card penetration rate among STO users	medium
Destination inference	heuristic approach	weak
Activity location inference	heuristic approach	medium
Research hypothesis	STO accessibility: walking distance threshold	medium
	living location inference	medium
Association part	deterministic link weights	medium

5.3 Future work

Source of errors can be categorized in a) errors related to data, b) errors related to hypotheses made and c) errors due to implementation of the methodology. A major limitation is that we used old data (both travel survey and smart card data) in a city where public transit modal share is small. Using the same methodology with recent data would help achieve better and more understandable results. Also, the alighting estimation module could be improved with latest work.

A major point would be to use data that we can verify at microscopic level as Munizaga et al. (2014) did. Building up a confidence in our methodology is the first step. Once this is done, we use it to emulate information about the transportation network between two travel surveys. This second step would also help validating the methodology since between two travel surveys, local on-board surveys are held for specific purposes and those surveys could help validate that our methodology is able to reproduce locally the right information through time.

This work is mainly about using passive data from AFC system, therefore we had information only about STO users. However it is not the only source of passive data that we could processed into a trip chain. Among data sources that we could use in a similar way we find social networks, bike sharing, car sharing etc. Even richer information could be used such as GPS data, streaming biking data; but these data sources could benefit from more work to extend our framework to them. Including more data sources would allow us to reach a larger population than just STO users.

REFERENCES

T. Adler and M. Ben-Akiva, "A theoretical and empirical model of trip chaining behavior," *Transportation Research Part B: Methodological*, vol. 13, no. 3, pp. 243–257, 1979.

B. Agard, C. Morency, and M. Trépanier, "Mining public transport user behaviour from smart card data," in *12th IFAC symposium on information control problems in manufacturing-INCOM*, 2006, pp. 17–19.

A. A. Alsger, M. Mesbah, L. Ferreira, and H. Safi, "Public transport origin-destination estimation using smart card fare data," in *Transportation Research Board 94th Annual Meeting*, no. 15-0801, 2015.

P. Anderson, B. Farooq, D. Efthymiou, and M. Bierlaire, "Associations generation in synthetic population for transportation applications: Graph-theoretic solution," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2429, pp. 38–50, 2014.

K. W. Axhausen, "Can we ever obtain the data we would like to have," *Theoretical foundations of travel choice modeling*, pp. 305–323, 1998.

K. W. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt, "Observing the rhythms of daily life: A six-week travel diary," *Transportation*, vol. 29, no. 2, pp. 95–124, 2002.

M. Bagchi and P. White, "What role for smart-card data from bus systems?" *Municipal Engineer*, vol. 157, no. 1, pp. 39–46, 2004.

——, "The potential of public transport smart card data," *Transport Policy*, vol. 12, no. 5, pp. 464–474, 2005.

D. Ballas, G. Clarke, and I. Turton, "Exploring microsimulation methodologies for the estimation of household attributes," in *4th International Conference on GeoComputation, Mary Washington College, Virginia, USA*, 1999.

P. Bass, P. Donoso, and M. Munizaga, "A model to assess public transport demand stability," *Transportation research part A: policy and practice*, vol. 45, no. 8, pp. 755–764, 2011.

C. Bayart and P. Bonnel, "L'impact du mode d'enquête sur la mesure des comportements de mobilité," *Economie et statistique*, vol. 437, no. 1, pp. 47–70, 2010.

——, "How to combine survey media (web, telephone, face-to-face): Lyon and rhône-alps case study," *Transportation Research Procedia*, vol. 11, pp. 118–135, 2015.

C. Bayart, P. Bonnel, C. Morency *et al.*, "Survey mode integration and data fusion: methods and challenges," *Transport Survey Methods: Keeping up with a Changing World*, pp. 587–611, 2009.

C. R. Bhat, "The maximum approximate composite marginal likelihood (macml) estimation of multinomial probit-based unordered response choice models," *Transportation Research Part B: Methodological*, vol. 45, no. 7, pp. 923–939, 2011.

C. R. Bhat and N. Eluru, "Recent advances in discrete and discrete-continuous modeling systems," *Feasibility decisions in transportation engineering,(ed. S. Nocera)*, pp. 111–144, 2010.

C. R. Bhat and S. Singh, "A joint model of work mode choice, evening commute stops and post-home arrival stops," *Final Report, Submitted to US DOT Region One, MIT*, 1997.

M. Bierlaire, "Biogeme: a free package for the estimation of discrete choice models," in *Swiss Transport Research Conference*, no. TRANSP-OR-CONF-2006-048, 2003.

M. Bierlaire, A. de Palma, R. Hurtubia, and P. Waddell, *Integrated Transport and Land Use Modeling for Sustainable Cities*. EPFL Press, 2015, no. EPFL-BOOK-207449.

C. Blanchette, *Analyse comparative entre les enquetes menages origine-destination et les systemes de paiement par carte a puce en transport urbain*. ProQuest, 2009.

P. Bonnel, "Prévision de la demande de transport," Ph.D. dissertation, Université Lumière-Lyon II, 2002.

J. L. Bowman and M. E. Ben-Akiva, "Activity-based disaggregate travel demand model system with activity schedules," *Transportation Research Part A: Policy and Practice*, vol. 35, no. 1, pp. 1–28, 2001.

S. Canada, "Canadian Census - fares and payments," https://www12.statcan.gc.ca, 2016, accessed: 2016-02-10.

A. Chakirov and A. Erath, "Use of public transport smart card fare payment data for travel behaviour analysis in singapore," 2011.

——, "Activity identification and primary location modelling based on smart card payment data for public transport," 2012.

K. Chan, "Asymptotic behavior of the gibbs sampler," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 320–326, 1993.

R. Chapleau, "Travel demand characterization for a large urban transit infrastructure using smart card transaction data," in *13th Word Conference on Transportation Research*, 2013.

R. Chapleau, M. Trépanier, and K. K. Chu, "The ultimate survey for transit planning: complete information with smart card data and gis," in *Proceedings of the 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability*, 2008, pp. 25–31.

K. K. Chu, "Leveraging data from a smart card automatic fare collection system for public transit planning," Ph.D. dissertation, École Polytechnique de Montréal, 2010.

S. Copsey, J. Sykes, J. Cecil, S. Walsh, N. Reed, J. Verity, S. Joseph, R. Southern, and A. Bygrave, "Implementing public transport smart mobile ticketing solutions within a deregulated shire environment–a united kingdom case study approach," in *European Transport Conference 2014*, 2014.

C. Cottrill, "Approaches to privacy preservation in intelligent transportation systems and vehicle-infrastructure integration initiative," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2129, pp. 9–15, 2009.

C.-l. Cui, Y.-l. Zhao, and Z.-y. Duan, "Research on the stability of public transit passenger travel behavior based on smart card data," in *CICTP 2014@ sSafe, Smart, and Sustainable Multimodal Transportation Systems*. ASCE, 2014, pp. 1318–1326.

J. de Dios Ortúzar and L. G. Willumsen, *Modelling transport*. John Wiley & Sons, 2011.

A. De Palma, M. De Lapparent, and N. Picard, "Modeling real estate investment decisions in households," 2014.

E. Deakin and S. Kim, "Transportation technologies: Implications for planning," 2001.

W. E. Deming and F. F. Stephan, "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known," *The Annals of Mathematical Statistics*, vol. 11, no. 4, pp. 427–444, 1940.

F. Devillaine, M. Munizaga, and M. Trépanier, "Detection of activities of public transport users by analyzing smart card data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2276, pp. 48–55, 2012.

S. Doherty and K. W. Axhausen, "The development of a unified modeling framework for the household activity-travel scheduling process," in *Traffic and Mobility*. Springer, 1999, pp. 35–56.

O. Egu, "Analyse du potentiel des données billettiques, le cas de lyon," Master's thesis, Ecole Nationale des Travaux Public de l'Etat, 2015.

N. Eluru, A. Pinjari, J. Guo, I. Sener, S. Srinivasan, R. Copperman, and C. Bhat, "Population updating system structures and models embedded in the comprehensive econometric microsimulator for urban systems," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2076, pp. 171–182, 2008.

G. D. Erhardt, "How smart is your smart card? evaluating transit smart card data with privacy restrictions and limited penetration rates," in *Transportation Research Board 95th Annual Meeting*, 2016.

B. Farooq, "Simpsinz," https://github.com/billjee/simpsynz, 2013, accessed: 2016-02-10.

B. Farooq, M. Bierlaire, R. Hurtubia, and G. Flötteröd, "Simulation based population synthesis," *Transportation Research Part B: Methodological*, vol. 58, pp. 243–263, 2013.

B. Farooq, R. Hurtubia, and M. Bierlaire, *Simulation based generation of the synthetic populations for Brussels cas study, Chapter 4: Case studies, in SustainCityHandook*. EPFL Press, 2013.

B. Farooq, E. J. Miller, F. Chingcuanco, and M. Giroux-Cook, "Microsimulation framework for urban price-taker markets," *Journal of Transport and Land Use*, vol. 6, no. 1, pp. 41–51, 2013.

B. Farooq, K. Muller, M. Bierlaire, and K. W. Axhausen, *Methodologies for synthesizing populations, Chapter 2: Modeling/Methodological contributions, in SustainCityHandook*. EPFL Press, 2013.

A. Fink, *How to Conduct Surveys: A Step-by-Step Guide: A Step-by-Step Guide*. New-York, USA: Sage Publications, 2012.

S. Fujii and R. Kitamura, "Evaluation of trip-inducing effects of new freeways using a structural equations model system of commuters' time use and travel," *Transportation Research Part B: Methodological*, vol. 34, no. 5, pp. 339–354, 2000.

P. Gaudette, "Microsimulation d'un réseau d'autobus défini dans le format gtfs," Master's thesis, Ecole Polytechnique de Montreal, 2015.

P. Gaudette, R. Chapleau, and T. Spurr, "Bus network microsimulation with gtfs and tap-in only smart card data," in *Transportation Research Board 95th Annual Meeting*, no. 16-5250, 2016.

C. J. Geyer, "Practical markov chain monte carlo," *Statistical Science*, pp. 473–483, 1992.

T. F. Golob, "A simultaneous model of household activity participation and trip chain generation," *Transportation Research Part B: Methodological*, vol. 34, no. 5, pp. 355–376, 2000.

——, "Structural equation modeling for travel behavior research," *Transportation Research Part B: Methodological*, vol. 37, no. 1, pp. 1–25, 2003.

A. Grapperon, "Bataclan," https://github.com/AntoineGrapperon/Bataclan, 2016, accessed: 2016-04-10.

J. Guo and C. R. Bhat, "Representation and analysis plan and data needs analysis for the activity-travel system," Center for Transportation Research, University of Texas at Austin, Tech. Rep., 2001.

T. Hägerstraand, "What about people in regional science?" *Papers in regional science*, vol. 24, no. 1, pp. 7–24, 1970.

L. He and M. Trépanier, "Estimating the destination of unlinked trips in public transportation smart card fare collection systems," in *Transportation Research Board 94th Annual Meeting*, no. 15-3433, 2015.

D. A. Hensher and T. Ton, "Tresis: A transportation, land use and environmental strategy impact simulator for urban areas," *Transportation*, vol. 29, no. 4, pp. 439–457, 2002.

S. Hess, M. Fowler, T. Adler, and A. Bahreinian, "A joint model for vehicle type and fuel type choice: evidence from a cross-nested logit study," *Transportation*, vol. 39, no. 3, pp. 593–625, 2012.

M. Hofmann, S. P. Wilson, and P. White, "Automated identification of linked trips at trip level using electronic fare collection data," in *Transportation Research Board 88th Annual Meeting*, no. 09-2417, 2009.

Z. Huang and P. Williamson, "A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata," *Department of Geography, University of Liverpool*, 2001.

iTRANS Consulting Inc., "Enquete origine-destination 2005," 2006.

L.-M. Kieu, A. Bhaskar, and E. Chung, "A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card afc data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 193–207, 2015.

T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 179–191, 2014.

M. Lovrić, T. Li, and P. Vervest, "Sustainable revenue management: A smart card enabled agent-based modeling approach," *Decision Support Systems*, vol. 54, no. 4, pp. 1587–1601, 2013.

M. D. Meyer and E. J. Miller, *Urban transportation planning: a decision-oriented approach*, 2001.

E. Miller and M. Roorda, "Prototype model of household activity-travel scheduling," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1831, pp. 114–121, 2003.

E. Miller, B. Farooq, F. Chingcuanco, and D. Wang, "Historical validation of integrated transport-land use model system," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2255, pp. 91–99, 2011.

E. J. Miller and P. A. Salvini, "The integrated land use, transportation, environment (ilute) microsimulation modelling system: description and current status," *Travel Behaviour Research: The Leading Edge*, pp. 711–724, 2001.

E. J. Miller, J. D. Hunt, J. E. Abraham, and P. A. Salvini, "Microsimulating urban systems," *Computers, environment and urban systems*, vol. 28, no. 1, pp. 9–44, 2004.

E. J. Miller, M. J. Roorda, and J. A. Carrasco, "A tour-based model of travel mode choice," *Transportation*, vol. 32, no. 4, pp. 399–422, 2005.

P. L. Mokhtarian, "Telecommuting and travel: state of the practice, state of the art," *Transportation*, vol. 18, no. 4, pp. 319–342, 1991.

C. Morency, M. Trepanier, and B. Agard, "Measuring transit use variability with smart-card data," *Transport Policy*, vol. 14, no. 3, pp. 193–203, 2007.

K. Müller and K. W. Axhausen, *Population synthesis for microsimulation: State of the art.* ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT), 2010.

M. Munizaga, F. Devillaine, C. Navarrete, and D. Silva, "Validating travel behavior estimated from smartcard data," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 70–79, 2014.

M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

N. Nassir, M. Hickman, and Z.-L. Ma, "Activity detection and transfer identification for public transit fare card data," *Transportation*, vol. 42, no. 4, pp. 683–705, 2015.

M. A. Ortega-Tong, "Classification of london's public transport users using smart card data," Ph.D. dissertation, Massachusetts Institute of Technology, 2013.

R. Paleti, R. M. Pendyala, C. R. Bhat, and K. C. Konduri, "A joint tour-based model of tour complexity, passenger accompaniment, vehicle type choice, and tour length," *Technical paper, Arizona State University, Tempe, AZ*, 2011.

M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

R. M. Pendyala, R. Kitamura, A. Kikuchi, T. Yamamoto, and S. Fujji, "Famos: The florida activity mobility simulator," in *Proceedings of the 84th annual meeting of the transportation research board*, 2005, pp. 9–13.

D. R. Pritchard, "Synthesizing agents and relationships for land use/transportation modelling," Ph.D. dissertation, University of Toronto, 2008.

F. Pukelsheim and B. Simeone, "On the iterative proportional fitting procedure: Structure of accumulation points and l1-error analysis," *Preprint*, 2009.

C. Z. Renner, T. W. Nicolai, and K. Nagel, "Agent-based land use transport interaction modeling: state of the art," 2014.

M. J. Roorda, S. T. Doherty, and E. J. Miller, "Operationalising household activity scheduling models: addressing assumptions and the use of new sources of behavioral data," *Integrated land-use and transportation models: Behavioural foundations*, pp. 61–85, 2005.

M. J. Roorda, J. A. Carrasco, and E. J. Miller, "An integrated model of vehicle transactions, activity scheduling and mode choice," *Transportation Research Part B: Methodological*, vol. 43, no. 2, pp. 217–229, 2009.

J. Ryan, H. Maoh, and P. Kanaroglou, "Population synthesis: Comparing the major techniques using a small, complete population of firms," *Geographical Analysis*, vol. 41, no. 2, pp. 181–203, 2009.

P. Salvini and E. J. Miller, "Ilute: An operational prototype of a comprehensive microsimulation model of urban systems," *Networks and Spatial Economics*, vol. 5, no. 2, pp. 217–234, 2005.

SBS, "SBS transit - fares overview," https://www.sbstransit.com, 2016, accessed: 2016-02-01.

T. Spurr, R. Chapleau, and D. Piché, "Use of subway smart card transactions for the discovery and partial correction of travel survey bias," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2405, pp. 57–67, 2014.

T. Spurr, A. Chu, R. Chapleau, and D. Piché, "A smart card transaction "travel diary" to assess the accuracy of the montréal household travel survey," *Transportation Research Procedia*, vol. 11, pp. 350–364, 2015.

P. R. Stopher and S. P. Greaves, "Household travel surveys: Where are we going?" *Transportation Research Part A: Policy and Practice*, vol. 41, no. 5, pp. 367–381, 2007.

TfL, "Transport For London - fares and payments," https://tfl.gov.uk, 2016, accessed: 2016-02-01.

M. Trépanier and C. Morency, "Assessing transit loyalty with smart card data," in *12th World Conference on Transport Research, July*, 2010, pp. 11–15.

M. Trépanier and R. Chapleau, "Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination," *Canadian Journal of Civil Engineering*, vol. 28, no. 1, pp. 48–58, 2001.

——, "Linking transit operational data to road network with a transportation object-oriented gis," *URISA Journal*, vol. 13, no. 2, pp. 23–30, 2001.

M. Trépanier, N. Tranchant, and R. Chapleau, "Individual trip destination estimation in a transit smart card automated fare collection system," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 1–14, 2007.

M. Trépanier, C. Morency, and B. Agard, "Calculation of transit performance measures using smartcard data," *Journal of Public Transportation*, vol. 12, no. 1, p. 5, 2009.

M. Trépanier, K. M. Habib, and C. Morency, "Are transit users loyal? revelations from a hazard model based on smart card data," *Canadian Journal of Civil Engineering*, vol. 39, no. 6, pp. 610–618, 2012.

M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1971, pp. 119–126, 2006.

M. Venigalla, "Household travel survey data fusion issues," in *Resource Paper, National Household Travel Survey Conference: Understanding Our Nation's Travel*, vol. 1, no. 2, 2004.

C. Viggiano, H. N. Koutsopoulos, J. Attanucci, and N. H. Wilson, "Inferring public transport access distance from smart card registration and transaction data," in *Transportation Research Board 95th Annual Meeting*, 2016.

D. Voas and P. Williamson, "An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata," *International Journal of Population Geography*, vol. 6, no. 5, pp. 349–366, 2000.

P. Vovsha and E. Petersen, "Model for person and household mobility attributes," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2132, pp. 95–105, 2009.

P. Waddell, "Urbansim: Modeling urban development for land use, transportation, and environmental planning," *Journal of the American Planning Association*, vol. 68, no. 3, pp. 297–314, 2002.

P. Wagner and M. Wegener, "Urban land use, transport and environment models: Experiences with an integrated microscopic approach," *disP-The Planning Review*, vol. 43, no. 170, pp. 45–56, 2007.

P. White, M. Bagchi, H. Bataille, and S. M. East, "The role of smartcard data in public transport," in *12th World Conference on Transport Research, Lisbon, Paper*, no. 1461, 2010.

F. Yasmin, C. Morency, and M. J. Roorda, "Assessment of spatial transferability of an activity-based model, tasha," *Transportation Research Part A: Policy and Practice*, vol. 78, pp. 200–213, 2015.

X. Ye, K. Konduri, R. M. Pendyala, B. Sana, and P. Waddell, "A methodology to match distributions of both household and person attributes in the generation of synthetic populations," in *88th Annual Meeting of the Transportation Research Board, Washington, DC*, 2009.

J. Zhao, A. Rahbee, and N. H. Wilson, "Estimating a rail passenger trip origin-destination matrix using automatic data collection systems," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 5, pp. 376–387, 2007.

C. Zhong, X. Huang, S. M. Arisona, G. Schmitt, and M. Batty, "Inferring building functions from a probabilistic model using public transportation data," *Computers, Environment and Urban Systems*, vol. 48, pp. 124–137, 2014.

APPENDIX A UML diagrams

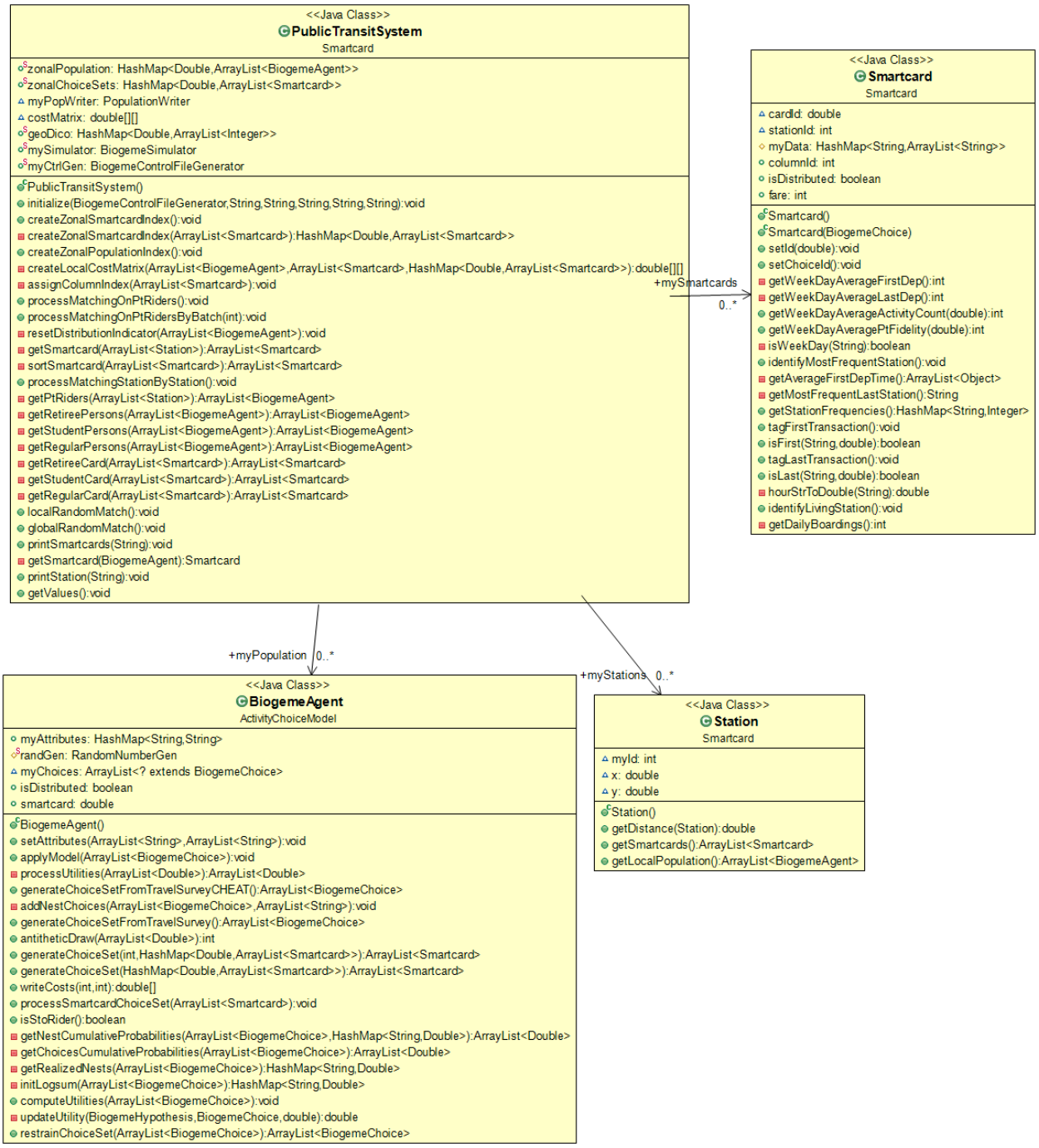

Figure A.1 UML diagram of the distribution of smart cards to synthesized agents.

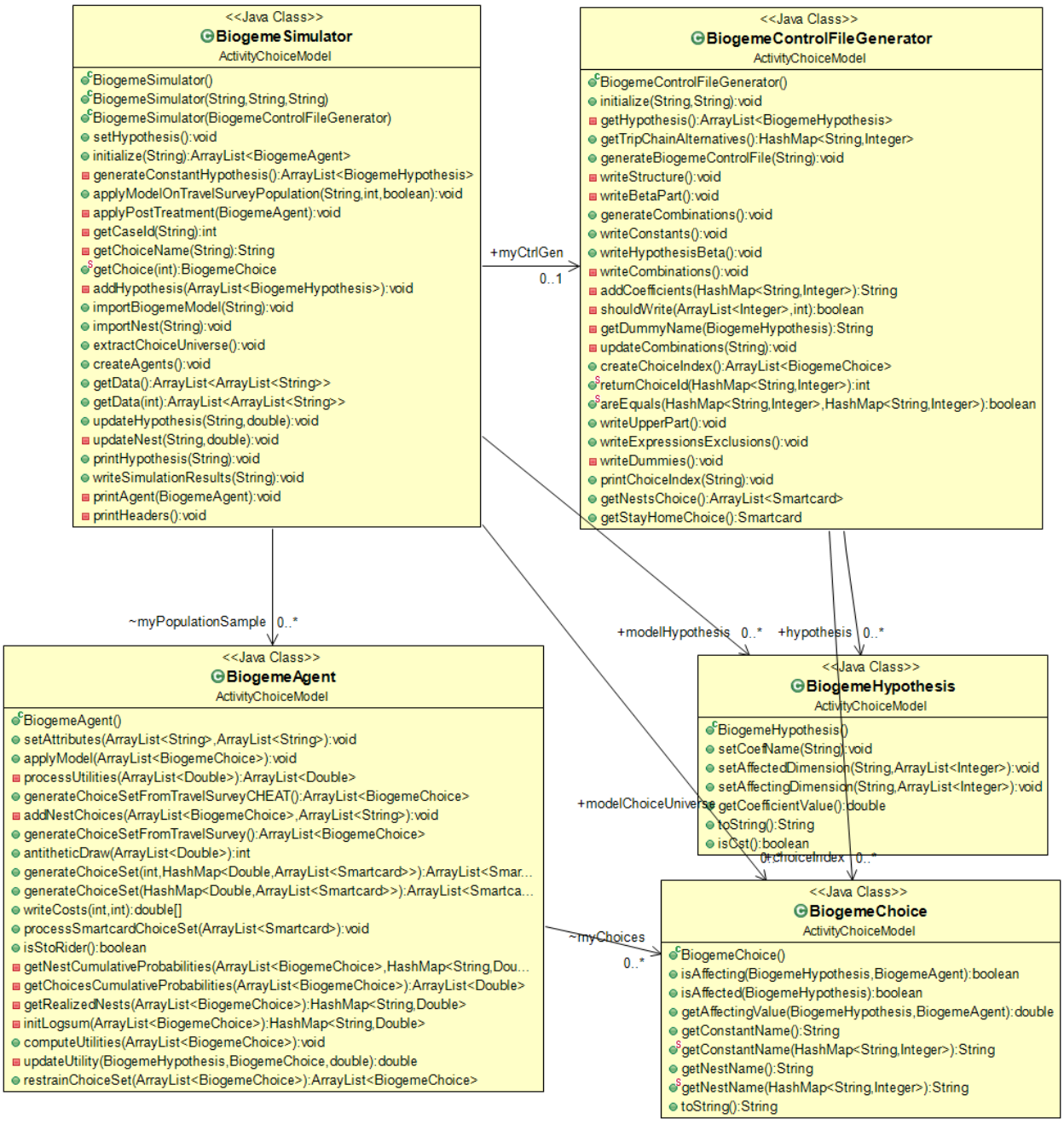

Figure A.2 UML diagram of the application of the trip chain choice model to agents.

APPENDIX B Full confusion matrix

Table B.1 – Complete confusion matrix

Observed choice	Simulated choice	Number of persons
C-1-0-1-0	C-2-1-1-1	22.04163
C-1-0-1-0	C-activeMode	65.87096013
C-1-0-1-0	C-carDriver	94.04035505
C-1-0-1-0	C-ptUserNoSto	53.28320172
C-1-0-1-1	C-1-0-2-2	16.40649609
C-1-0-1-1	C-1-1-1-1	18.26234671
C-1-0-1-1	C-1-1-1-2	13.25527438
C-1-0-1-1	C-1-1-2-2	33.37898268
C-1-0-1-1	C-1-2-3-2	14.68434775
C-1-0-1-1	C-2-0-1-1	18.95086765
C-1-0-1-1	C-2-1-1-1	75.67650928
C-1-0-1-1	C-activeMode	141.1742841
C-1-0-1-1	C-carDriver	674.1554681
C-1-0-1-1	C-carPassenger	232.4899935
C-1-0-1-1	C-ptUserNoSto	46.31901468
C-1-0-1-2	C-activeMode	16.54537394
C-1-0-1-2	C-carDriver	82.04764961
C-1-0-1-2	C-ptUserNoSto	38.28770232
C-1-0-2-0	C-carDriver	13.40806801
C-1-0-2-1	C-1-1-3-2	12.99674523
C-1-0-2-1	C-2-0-1-2	14.42411534
C-1-0-2-1	C-2-1-1-1	15.11100346
C-1-0-2-1	C-activeMode	120.624568
C-1-0-2-1	C-carDriver	368.1252101
C-1-0-2-1	C-carPassenger	35.05266196
C-1-0-2-2	C-1-0-1-1	36.51261342
C-1-0-2-2	C-1-1-1-1	16.37481972
C-1-0-2-2	C-1-1-2-2	22.61455974
C-1-0-2-2	C-1-1-3-2	19.55410862

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-1-0-2-2	C-1-2-2-1	14.69857295
C-1-0-2-2	C-2-0-1-0	16.25316435
C-1-0-2-2	C-2-1-1-1	28.07778958
C-1-0-2-2	C-2-2-1-1	18.00472723
C-1-0-2-2	C-activeMode	45.84947165
C-1-0-2-2	C-carDriver	581.9463018
C-1-0-2-2	C-carPassenger	57.95780729
C-1-0-2-2	C-ptUserNoSto	120.4034409
C-1-0-3-0	C-activeMode	17.81668239
C-1-0-3-0	C-carDriver	34.78650379
C-1-0-3-1	C-1-1-2-2	16.79030133
C-1-0-3-1	C-activeMode	19.42375508
C-1-0-3-1	C-carDriver	63.02928945
C-1-0-3-1	C-carPassenger	46.3704341
C-1-0-3-2	C-1-0-2-1	24.97872412
C-1-0-3-2	C-1-0-2-2	17.79939345
C-1-0-3-2	C-1-1-1-1	13.37053764
C-1-0-3-2	C-1-1-2-1	18.09167057
C-1-0-3-2	C-2-0-1-0	17.54342192
C-1-0-3-2	C-2-0-1-1	20.68230571
C-1-0-3-2	C-2-1-1-0	22.34753377
C-1-0-3-2	C-2-2-1-2	17.13385038
C-1-0-3-2	C-activeMode	34.46564788
C-1-0-3-2	C-carDriver	541.4726758
C-1-0-3-2	C-carPassenger	157.7414213
C-1-0-3-2	C-ptUserNoSto	65.20176195
C-1-1-1-0	C-1-0-2-2	22.04163
C-1-1-1-0	C-1-1-3-2	46.1723544
C-1-1-1-0	C-1-2-1-2	10.75072075
C-1-1-1-0	C-2-0-2-1	10.75072075
C-1-1-1-0	C-2-1-1-1	18.61777926
C-1-1-1-0	C-activeMode	65.88676927
C-1-1-1-0	C-carDriver	100.2255657

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-1-1-1-0	C-carPassenger	35.21572984
C-1-1-1-0	C-ptUserNoSto	72.9169797
C-1-1-1-1	C-1-0-1-1	15.11100346
C-1-1-1-1	C-1-0-2-2	77.45807773
C-1-1-1-1	C-1-1-1-1	14.66521781
C-1-1-1-1	C-1-1-2-2	29.31912945
C-1-1-1-1	C-1-1-3-2	27.75858515
C-1-1-1-1	C-1-2-1-2	35.50059828
C-1-1-1-1	C-2-0-1-1	21.3454417
C-1-1-1-1	C-2-1-1-1	183.667785
C-1-1-1-1	C-activeMode	320.6774106
C-1-1-1-1	C-carDriver	1567.565101
C-1-1-1-1	C-carPassenger	256.4077358
C-1-1-1-1	C-ptUserNoSto	155.4809867
C-1-1-1-2	C-1-0-2-2	14.70177878
C-1-1-1-2	C-1-1-1-1	28.72298742
C-1-1-1-2	C-1-2-1-2	18.79943503
C-1-1-1-2	C-2-1-1-1	33.15945512
C-1-1-1-2	C-2-2-1-1	16.78512911
C-1-1-1-2	C-activeMode	36.10744177
C-1-1-1-2	C-carDriver	204.4155119
C-1-1-1-2	C-carPassenger	46.48019496
C-1-1-1-2	C-ptUserNoSto	58.97952898
C-1-1-2-0	C-1-1-2-2	16.13493408
C-1-1-2-0	C-carDriver	13.19288865
C-1-1-2-0	C-carPassenger	16.93914631
C-1-1-2-0	C-ptUserNoSto	13.67441187
C-1-1-2-1	C-1-0-1-1	19.78732135
C-1-1-2-1	C-1-1-1-1	29.27023136
C-1-1-2-1	C-1-1-2-1	15.0488175
C-1-1-2-1	C-1-1-2-2	39.77257973
C-1-1-2-1	C-1-2-1-1	20.9385442
C-1-1-2-1	C-2-1-1-1	13.31263433

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-1-1-2-1	C-activeMode	142.2700061
C-1-1-2-1	C-carDriver	559.8332696
C-1-1-2-1	C-carPassenger	58.22234359
C-1-1-2-1	C-ptUserNoSto	25.82460943
C-1-1-2-2	C-1-0-1-1	42.97821884
C-1-1-2-2	C-1-0-2-1	14.63503588
C-1-1-2-2	C-1-0-2-2	13.67441187
C-1-1-2-2	C-1-0-3-2	31.44153665
C-1-1-2-2	C-1-1-1-1	12.97519194
C-1-1-2-2	C-1-1-1-2	23.12138307
C-1-1-2-2	C-1-1-2-2	57.61366703
C-1-1-2-2	C-1-1-3-1	46.73064555
C-1-1-2-2	C-1-1-3-2	27.96137706
C-1-1-2-2	C-1-2-2-2	12.40009539
C-1-1-2-2	C-1-2-3-2	21.80905883
C-1-1-2-2	C-2-0-1-1	16.52952466
C-1-1-2-2	C-2-1-1-1	145.1945618
C-1-1-2-2	C-2-2-1-2	16.14421592
C-1-1-2-2	C-activeMode	354.7971407
C-1-1-2-2	C-carDriver	1586.760407
C-1-1-2-2	C-carPassenger	454.1839254
C-1-1-2-2	C-ptUserNoSto	392.0400325
C-1-1-3-0	C-activeMode	23.14197143
C-1-1-3-0	C-carDriver	9.547890651
C-1-1-3-0	C-carPassenger	30.9188818
C-1-1-3-1	C-1-1-3-1	12.64699232
C-1-1-3-1	C-1-2-1-1	13.79475383
C-1-1-3-1	C-2-0-2-0	13.63523093
C-1-1-3-1	C-2-1-1-1	89.72861597
C-1-1-3-1	C-2-1-1-2	20.81016541
C-1-1-3-1	C-activeMode	33.67371868
C-1-1-3-1	C-carDriver	322.2511559
C-1-1-3-1	C-carPassenger	87.48325846

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-1-1-3-1	C-ptUserNoSto	26.73184742
C-1-1-3-2	C-1-0-2-2	17.54342192
C-1-1-3-2	C-1-1-1-2	16.78512911
C-1-1-3-2	C-1-1-2-1	36.21702814
C-1-1-3-2	C-2-0-1-1	13.67441187
C-1-1-3-2	C-2-1-1-0	13.37339486
C-1-1-3-2	C-2-1-1-1	80.14765461
C-1-1-3-2	C-2-2-1-1	11.87198855
C-1-1-3-2	C-activeMode	200.9722015
C-1-1-3-2	C-carDriver	1021.911049
C-1-1-3-2	C-carPassenger	307.43806
C-1-1-3-2	C-ptUserNoSto	115.7586691
C-1-2-1-0	C-1-1-3-2	14.63503588
C-1-2-1-0	C-2-0-1-1	16.97372093
C-1-2-1-0	C-2-1-1-1	36.97707094
C-1-2-1-0	C-activeMode	17.4424626
C-1-2-1-0	C-carDriver	50.97171905
C-1-2-1-0	C-carPassenger	27.96393039
C-1-2-1-0	C-ptUserNoSto	19.90968652
C-1-2-1-1	C-1-0-2-1	26.73184742
C-1-2-1-1	C-1-0-3-2	13.82037488
C-1-2-1-1	C-1-1-1-0	13.63471419
C-1-2-1-1	C-1-1-2-1	13.37515283
C-1-2-1-1	C-1-1-3-2	16.77334903
C-1-2-1-1	C-activeMode	89.74104418
C-1-2-1-1	C-carDriver	111.71746
C-1-2-1-1	C-carPassenger	81.72109521
C-1-2-1-1	C-ptUserNoSto	106.2833898
C-1-2-1-2	C-1-1-2-2	9.545975305
C-1-2-1-2	C-1-2-2-2	30.94195282
C-1-2-1-2	C-activeMode	83.57018627
C-1-2-1-2	C-carDriver	255.3686504
C-1-2-1-2	C-carPassenger	126.1784095

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-1-2-1-2	C-ptUserNoSto	27.84246914
C-1-2-2-0	C-carDriver	34.9498735
C-1-2-2-1	C-activeMode	91.70299516
C-1-2-2-1	C-carDriver	31.9986928
C-1-2-2-1	C-carPassenger	42.95942528
C-1-2-2-1	C-ptUserNoSto	40.68856458
C-1-2-2-2	C-1-1-2-2	36.0948907
C-1-2-2-2	C-2-1-1-1	53.13567786
C-1-2-2-2	C-activeMode	54.04442447
C-1-2-2-2	C-carDriver	203.9073437
C-1-2-2-2	C-carPassenger	110.3155445
C-1-2-2-2	C-ptUserNoSto	60.1463658
C-1-2-3-0	C-2-1-1-1	15.88272873
C-1-2-3-0	C-activeMode	21.97239089
C-1-2-3-0	C-carDriver	20.3882306
C-1-2-3-0	C-carPassenger	30.9188818
C-1-2-3-2	C-1-0-2-1	27.63855047
C-1-2-3-2	C-2-0-1-1	14.02489377
C-1-2-3-2	C-2-1-1-1	75.40131326
C-1-2-3-2	C-activeMode	32.94361432
C-1-2-3-2	C-carDriver	47.79314045
C-1-2-3-2	C-carPassenger	25.79506255
C-1-2-3-2	C-ptUserNoSto	14.16805727
C-2-0-1-0	C-2-1-1-1	9.706126242
C-2-0-1-0	C-2-1-1-2	14.77498101
C-2-0-1-0	C-2-2-2-1	30.6063791
C-2-0-1-0	C-activeMode	38.24725829
C-2-0-1-0	C-carDriver	274.9575368
C-2-0-1-0	C-carPassenger	51.79507901
C-2-0-1-0	C-ptUserNoSto	63.23689341
C-2-0-1-1	C-1-1-1-1	37.4184054
C-2-0-1-1	C-1-1-2-2	13.56673699
C-2-0-1-1	C-1-2-3-2	16.55789505

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-2-0-1-1	C-2-1-1-1	51.60680584
C-2-0-1-1	C-2-1-2-2	15.88272873
C-2-0-1-1	C-activeMode	122.6856474
C-2-0-1-1	C-carDriver	810.7049108
C-2-0-1-1	C-carPassenger	188.8156604
C-2-0-1-1	C-ptUserNoSto	81.00290376
C-2-0-1-2	C-carPassenger	44.12979878
C-2-0-2-0	C-2-0-1-1	17.02477414
C-2-0-2-1	C-activeMode	27.65009986
C-2-0-2-1	C-carDriver	19.97734079
C-2-0-2-2	C-1-2-1-2	21.2108575
C-2-1-1-0	C-1-0-1-1	13.37339486
C-2-1-1-0	C-1-0-2-1	18.42645909
C-2-1-1-0	C-1-1-1-0	13.82037488
C-2-1-1-0	C-1-1-2-1	12.44711328
C-2-1-1-0	C-1-1-2-2	15.90318262
C-2-1-1-0	C-1-1-3-2	14.16805727
C-2-1-1-0	C-1-2-1-2	16.97372093
C-2-1-1-0	C-1-2-3-2	12.94667302
C-2-1-1-0	C-2-0-1-1	11.33006465
C-2-1-1-0	C-2-1-1-0	13.63471419
C-2-1-1-0	C-2-1-1-1	45.9507226
C-2-1-1-0	C-2-2-1-1	17.8112817
C-2-1-1-0	C-activeMode	232.0333396
C-2-1-1-0	C-carDriver	301.981293
C-2-1-1-0	C-carPassenger	361.9980762
C-2-1-1-0	C-ptUserNoSto	375.311918
C-2-1-1-1	C-1-0-1-1	100.2326943
C-2-1-1-1	C-1-0-2-1	32.25099613
C-2-1-1-1	C-1-0-2-2	60.56359417
C-2-1-1-1	C-1-0-3-2	46.34655499
C-2-1-1-1	C-1-1-1-0	13.79331979
C-2-1-1-1	C-1-1-1-1	55.61419505

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-2-1-1-1	C-1-1-1-2	30.45202999
C-2-1-1-1	C-1-1-2-1	45.96881905
C-2-1-1-1	C-1-1-2-2	66.6271031
C-2-1-1-1	C-1-1-3-1	26.14779482
C-2-1-1-1	C-1-1-3-2	11.2959743
C-2-1-1-1	C-1-2-1-1	10.93553935
C-2-1-1-1	C-1-2-1-2	43.83466839
C-2-1-1-1	C-1-2-2-1	9.545975305
C-2-1-1-1	C-1-2-2-2	31.0948281
C-2-1-1-1	C-2-0-1-1	86.43262287
C-2-1-1-1	C-2-1-1-0	81.61492271
C-2-1-1-1	C-2-1-1-1	313.6272508
C-2-1-1-1	C-2-1-2-1	15.51105101
C-2-1-1-1	C-2-1-2-2	10.93553935
C-2-1-1-1	C-2-1-3-2	11.7861243
C-2-1-1-1	C-2-2-1-0	15.17634986
C-2-1-1-1	C-2-2-1-1	11.7861243
C-2-1-1-1	C-2-2-1-2	16.98304201
C-2-1-1-1	C-activeMode	898.3744847
C-2-1-1-1	C-carDriver	2329.415883
C-2-1-1-1	C-carPassenger	640.4324557
C-2-1-1-1	C-ptUserNoSto	771.5404289
C-2-1-1-2	C-1-1-1-1	11.14039167
C-2-1-1-2	C-1-1-3-1	12.49001399
C-2-1-1-2	C-1-2-1-1	17.30029887
C-2-1-1-2	C-2-0-1-1	11.14039167
C-2-1-1-2	C-2-1-1-1	30.1448612
C-2-1-1-2	C-2-2-1-1	31.38304095
C-2-1-1-2	C-carDriver	191.8735938
C-2-1-1-2	C-carPassenger	11.33006465
C-2-1-1-2	C-ptUserNoSto	14.57503931
C-2-1-2-0	C-1-0-3-2	24.60357589
C-2-1-2-0	C-carDriver	12.26944763

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-2-1-2-1	C-activeMode	22.3294655
C-2-1-2-1	C-carDriver	54.70706463
C-2-1-2-1	C-carPassenger	12.19680812
C-2-1-2-2	C-1-0-3-1	16.37481972
C-2-1-2-2	C-1-1-2-1	12.44711328
C-2-1-2-2	C-activeMode	13.63471419
C-2-1-2-2	C-carDriver	45.19785917
C-2-1-2-2	C-carPassenger	16.34930011
C-2-1-2-2	C-ptUserNoSto	75.02895328
C-2-1-3-2	C-ptUserNoSto	13.37339486
C-2-2-1-0	C-1-0-1-1	14.00009341
C-2-2-1-0	C-1-1-1-1	54.75438913
C-2-2-1-0	C-1-1-2-2	30.89935543
C-2-2-1-0	C-1-1-3-2	13.67441187
C-2-2-1-0	C-2-1-1-1	53.27567735
C-2-2-1-0	C-activeMode	178.0918676
C-2-2-1-0	C-carDriver	122.116359
C-2-2-1-0	C-carPassenger	105.8822938
C-2-2-1-0	C-ptUserNoSto	57.21214468
C-2-2-1-1	C-1-0-1-1	18.00472723
C-2-2-1-1	C-1-0-1-2	35.40566325
C-2-2-1-1	C-1-0-2-2	18.5271652
C-2-2-1-1	C-1-1-1-1	26.97800474
C-2-2-1-1	C-1-1-3-2	17.8112817
C-2-2-1-1	C-2-0-1-1	13.6668283
C-2-2-1-1	C-2-1-1-1	65.19324842
C-2-2-1-1	C-2-2-1-1	27.47203783
C-2-2-1-1	C-activeMode	175.1842571
C-2-2-1-1	C-carDriver	175.8809016
C-2-2-1-1	C-carPassenger	173.9257324
C-2-2-1-1	C-ptUserNoSto	149.6989502
C-2-2-1-2	C-1-0-1-1	20.98963787
C-2-2-1-2	C-1-0-3-2	20.90763363

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-2-2-1-2	C-1-2-2-2	24.5301226
C-2-2-1-2	C-2-0-1-1	35.6388757
C-2-2-1-2	C-2-1-1-1	36.70944983
C-2-2-1-2	C-activeMode	114.0287215
C-2-2-1-2	C-carDriver	213.1165272
C-2-2-1-2	C-carPassenger	107.1726325
C-2-2-1-2	C-ptUserNoSto	141.7567069
C-2-2-2-0	C-1-0-3-2	17.4424626
C-2-2-2-0	C-ptUserNoSto	23.03207845
C-2-2-2-1	C-2-2-1-0	32.88066558
C-2-2-2-1	C-carPassenger	22.4880136
C-2-2-2-2	C-2-0-1-1	21.08862013
C-2-2-2-2	C-carDriver	47.51470146
C-activeMode	C-1-0-1-1	138.9479735
C-activeMode	C-1-0-2-1	94.26144342
C-activeMode	C-1-0-2-2	80.44946282
C-activeMode	C-1-0-3-0	29.44791836
C-activeMode	C-1-0-3-1	58.97651041
C-activeMode	C-1-0-3-2	252.8604641
C-activeMode	C-1-1-1-0	48.810214
C-activeMode	C-1-1-1-1	320.7429948
C-activeMode	C-1-1-1-2	67.41267732
C-activeMode	C-1-1-2-1	74.76926591
C-activeMode	C-1-1-2-2	383.4407491
C-activeMode	C-1-1-3-0	44.79778248
C-activeMode	C-1-1-3-1	137.4854909
C-activeMode	C-1-1-3-2	186.9089419
C-activeMode	C-1-2-1-2	67.12440362
C-activeMode	C-1-2-2-1	15.27867521
C-activeMode	C-1-2-2-2	74.85263583
C-activeMode	C-1-2-3-2	22.15973937
C-activeMode	C-2-0-1-0	57.81103606
C-activeMode	C-2-0-1-1	320.0490037

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-activeMode	C-2-0-1-2	47.4656405
C-activeMode	C-2-0-2-1	16.86649734
C-activeMode	C-2-1-1-0	164.9307327
C-activeMode	C-2-1-1-1	1155.338282
C-activeMode	C-2-2-1-0	52.10194096
C-activeMode	C-2-2-1-1	96.60616052
C-activeMode	C-2-2-1-2	81.69375654
C-activeMode	C-2-2-2-0	20.66330525
C-activeMode	C-activeMode	3043.041106
C-activeMode	C-carDriver	6472.947665
C-activeMode	C-carPassenger	2913.558981
C-activeMode	C-ptUserNoSto	2090.771868
C-carDriver	C-1-0-1-0	67.73265769
C-carDriver	C-1-0-1-1	783.6267634
C-carDriver	C-1-0-1-2	106.8094277
C-carDriver	C-1-0-2-0	11.99883298
C-carDriver	C-1-0-2-1	281.6931988
C-carDriver	C-1-0-2-2	367.5556368
C-carDriver	C-1-0-3-0	18.95086765
C-carDriver	C-1-0-3-1	58.77668993
C-carDriver	C-1-0-3-2	489.9889133
C-carDriver	C-1-1-1-0	153.6275849
C-carDriver	C-1-1-1-1	1212.356423
C-carDriver	C-1-1-1-2	231.169677
C-carDriver	C-1-1-2-0	23.66959126
C-carDriver	C-1-1-2-1	397.0639789
C-carDriver	C-1-1-2-2	1263.247853
C-carDriver	C-1-1-3-0	20.81016541
C-carDriver	C-1-1-3-1	388.9515375
C-carDriver	C-1-1-3-2	835.7694431
C-carDriver	C-1-2-1-0	12.99674523
C-carDriver	C-1-2-1-1	168.2743604
C-carDriver	C-1-2-1-2	288.8809638

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-carDriver	C-1-2-2-1	60.19745406
C-carDriver	C-1-2-2-2	224.0802181
C-carDriver	C-1-2-3-0	31.345197
C-carDriver	C-1-2-3-2	68.72322355
C-carDriver	C-2-0-1-0	114.843364
C-carDriver	C-2-0-1-1	951.8284752
C-carDriver	C-2-0-1-2	37.56442467
C-carDriver	C-2-0-2-2	15.11100346
C-carDriver	C-2-1-1-0	589.7965017
C-carDriver	C-2-1-1-1	4004.846927
C-carDriver	C-2-1-1-2	292.7940555
C-carDriver	C-2-1-2-0	13.4249812
C-carDriver	C-2-1-2-1	64.53705218
C-carDriver	C-2-1-2-2	104.0950378
C-carDriver	C-2-2-1-0	120.9179637
C-carDriver	C-2-2-1-1	527.8154716
C-carDriver	C-2-2-1-2	369.0576966
C-carDriver	C-2-2-2-2	55.1078898
C-carDriver	C-activeMode	9273.730128
C-carDriver	C-carDriver	88197.96284
C-carDriver	C-carPassenger	10756.46516
C-carDriver	C-ptUserNoSto	5115.782275
C-carPassenger	C-1-0-1-0	49.79051096
C-carPassenger	C-1-0-1-1	154.4027811
C-carPassenger	C-1-0-1-2	22.21980883
C-carPassenger	C-1-0-2-1	44.28690383
C-carPassenger	C-1-0-2-2	146.3924889
C-carPassenger	C-1-0-3-2	47.02889268
C-carPassenger	C-1-1-1-0	40.07273573
C-carPassenger	C-1-1-1-1	221.6442567
C-carPassenger	C-1-1-1-2	29.15177637
C-carPassenger	C-1-1-2-1	129.9552355
C-carPassenger	C-1-1-2-2	285.8926734

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-carPassenger	C-1-1-3-1	133.9283306
C-carPassenger	C-1-1-3-2	180.4182381
C-carPassenger	C-1-2-1-1	45.73371811
C-carPassenger	C-1-2-1-2	51.61755432
C-carPassenger	C-1-2-2-1	60.17285857
C-carPassenger	C-1-2-2-2	78.89158976
C-carPassenger	C-2-0-1-0	49.13061611
C-carPassenger	C-2-0-1-1	248.3954123
C-carPassenger	C-2-1-1-0	141.7755681
C-carPassenger	C-2-1-1-1	896.5148223
C-carPassenger	C-2-1-1-2	67.52098991
C-carPassenger	C-2-1-2-1	14.86400501
C-carPassenger	C-2-1-2-2	22.44639172
C-carPassenger	C-2-2-1-0	30.69450029
C-carPassenger	C-2-2-1-1	136.0595608
C-carPassenger	C-2-2-1-2	90.69608919
C-carPassenger	C-2-2-2-2	32.11813802
C-carPassenger	C-activeMode	2446.877757
C-carPassenger	C-carDriver	11980.84837
C-carPassenger	C-carPassenger	2926.72004
C-carPassenger	C-ptUserNoSto	2259.653195
C-ptUserNoSto	C-1-0-1-0	66.31140105
C-ptUserNoSto	C-1-0-1-1	324.9056655
C-ptUserNoSto	C-1-0-1-2	9.622089645
C-ptUserNoSto	C-1-0-2-2	83.73726133
C-ptUserNoSto	C-1-0-3-0	18.9979792
C-ptUserNoSto	C-1-0-3-1	62.07196449
C-ptUserNoSto	C-1-0-3-2	92.53162378
C-ptUserNoSto	C-1-1-1-0	16.89317439
C-ptUserNoSto	C-1-1-1-1	460.661814
C-ptUserNoSto	C-1-1-1-2	93.95875866
C-ptUserNoSto	C-1-1-2-1	198.6835937
C-ptUserNoSto	C-1-1-2-2	373.3322819

Continued on next page

Table B.1 Continued from previous page

Observed choice	Simulated choice	Number of persons
C-ptUserNoSto	C-1-1-3-1	43.29534709
C-ptUserNoSto	C-1-1-3-2	238.1363703
C-ptUserNoSto	C-1-2-1-1	29.45907808
C-ptUserNoSto	C-1-2-1-2	41.0955703
C-ptUserNoSto	C-1-2-2-1	41.71477002
C-ptUserNoSto	C-1-2-2-2	40.42636202
C-ptUserNoSto	C-1-2-3-0	38.14663931
C-ptUserNoSto	C-1-2-3-2	15.6325652
C-ptUserNoSto	C-2-0-1-0	57.86050347
C-ptUserNoSto	C-2-0-1-1	315.0075973
C-ptUserNoSto	C-2-1-1-0	222.1962149
C-ptUserNoSto	C-2-1-1-1	1007.922827
C-ptUserNoSto	C-2-1-1-2	31.90013964
C-ptUserNoSto	C-2-2-1-1	160.8903711
C-ptUserNoSto	C-2-2-1-2	92.73286449
C-ptUserNoSto	C-activeMode	3047.133506
C-ptUserNoSto	C-carDriver	2924.761748
C-ptUserNoSto	C-carPassenger	3555.568843
C-ptUserNoSto	C-ptUserNoSto	3423.59978

End of table