CIRRELT

Centre interuniversitaire de recherche
sur les réseaux d'entreprise, la logistique et le transport

Interuniversity Research Centre
on Enterprise Networks, Logistics and Transportation

Information Fusion of Smart Card Data with Travel Survey

Antoine Grapperon
Bilal Farooq
Martin Trépanier

October 2016

CIRRELT-2016-59

Bureaux de Montréal :
Université de Montréal
Pavillon André-Aisenstadt
C.P. 6128, succursale Centre-ville
Montréal (Québec)
Canada H3C 3J7
Téléphone : 514 343-7575
Télécopie : 514 343-7121

Bureaux de Québec :
Université Laval
Pavillon Palasis-Prince
2325, de la Terrasse, bureau 2642
Québec (Québec)
Canada G1V 0A6
Téléphone : 418 656-2073
Télécopie : 418 656-2624

www.cirrelt.ca

Information Fusion of Smart Card Data with Travel Survey

Antoine Grapperon[1,2,*], Bilal Farooq[1,2], Martin Trépanier[1,3]

[1] Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT)

[2] Department of Civil, Geological and Mining Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

[3] Department of Mathematical and Industrial Engineering, Polytechnique Montréal, P.O. Box 6079, Station Centre-ville, Montréal, Canada H3C 3A7

Abstract. In comparison to traditional travel surveys, smart card transaction data with extensive spatio-temporal richness offers a great opportunity. However the smart card dataset may lack some basic information such as trip destinations and socio-demographic attributes of the traveler. Our goal is to enrich smart card data with this missing information. We propose an information fusion approach of census, travel survey, and smart card data. Our methodology involves solving a maximum weighted bipartite graph consisting of a collection of smart card owners and a collection of smart cards. The links between the nodes are weighted using a trip chain choice model. This methodology is applied on Gatineau's public transit service (Canada). Results are analyzed at macroscopic and mesoscopic level by comparing estimated public transit riders' population with the observed population in the travel survey. Our method produces results that fit the expectation as high as 93%.

Keywords: Smart card, automated fare collection, public transportation, behavior, data fusion, information fusion, census, travel survey.

Acknowledgements. The authors wish to thanks Société des Transports de l'Outaouais for granting us access to their data. The authors are grateful to Natural Sciences and Engineering Research Council of Canada (NSERC RDCPJ 446 107-1) and Thalès for funding the project.

---

* Corresponding author: Antoine.Grapperon@cirrelt.ca

Dépôt légal – Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada, 2016

1 Introduction

The classic way to collect data for strategic and operational planning is through travel surveys and on-board surveys. They are interesting data sets since they include comprehensive socio-demographic information, they are designed to fulfill most of the data needs and often the survey methodology is not altered between two survey campaigns, which ensures data consistency over time. However, both travel and on-board surveys present various drawbacks. Response rates are decreasing, sampling strategies and theoretical background have to adapt to a more complex world (web based surveys, surveys assisted with GPS, multi-day surveys etc.) and their cost is increasing (Stopher and Greaves, 2007; Fink, 2012; de Dios Ortúzar and Willumsen, 2011; Bayart and Bonnel, 2015). In addition, the data they are producing do not answer every question, especially regarding longitudinal behaviours (Axhausen et al., 2002; Yáñez et al., 2010). Axhausen (1998) states that surveying is not likely to achieve the satisfactory level of detail required by transportation modellers nor a satisfactory sampling rate and response rate.

Automated fare collection (AFC) systems such as smart cards are now being used in many different cities. The AFC system gathers a massive volume of transaction information and recently, making sense out of this data has become an important research topic. It is an interesting dataset for four main reasons:

- it is not a sample since all the public transit riders are recorded (excepted free-riders)(Bagchi and White, 2005)
- it is passively collected data therefore there are no reporting biases. Although there could be a systematic bias due to hardware failures, actively using smart card data can help detect and fix them (Chapleau and Chu, 2007).
- it is longitudinal data: it captures the temporal evolution of riders' behaviour
- it is tied to detailed transit operational data (Bagchi and White, 2004), (Chapleau et al., 2008).

In summary, planners are interested in accessibility analysis of the public transit system, especially by cohorts of population, by location, by line etc. Conventional data (travel surveys) have the socio-demographics while smart card data have detailed longitudinal mobility information. The information from the two sources needs to be fused to develop detailed accessibility analysis.

The smart card data and travel survey data exist in four dimensional space: space, time, socio-demographic characteristics and mobility choices. The smart card data have more accurate and detailed information on space, time and mobility choices dimensions while lacking of socio-demographic information and activity purposes of trips (see Table 1). Making the best use of AFC datasets requires enriching them with socio-demographic information (Pelletier et al., 2011) to help understand individual's mobility (Bayart et al., 2009). Due to privacy concerns, no socio-demographic characteristics can be directly attached to smart card data (Cottrill, 2009). In some AFC system, the pricing policy can be very detailed and it can offer a good granularity to attach some information (usually: adult, student, senior). To the best of our knowledge, the only work that was done to attach socio-demographic attributes to smart cards can be found in Trépanier et al. (2012) and it was done at a macroscopic level. No disaggregate information could be drawn.

We propose a methodology to infer socio-demographic attributes associated to smart cards mainly based on the analysis of the spatio-temporal patterns in the smart card data. We choose this approach confidently since destination inferences methodologies are rather reliable (Munizaga et al., 2014) and smart card allows making a rather detailed analysis of public transit riders (Ortega-Tong, 2013). The methodology consists in defining a Maximum Weighted Bipartite matching problem.

The first part of the graph is composed of the population (potential smart card owners) and the second part of the graph is composed of the smart cards. We weight the links by applying a trip chain choice model on the population with a choice set constituted of trip chains observed through the AFC system. We apply the methodology to the Gatineau (Canada) case study and proceed to a partial validation.

The paper layout is as follows. The current and first section introduces the paper. The second section is a literature review of smart card data enrichment. The third section describes our methodology: first the context is described, secondly the theoretical description of the methodology is given, the fourth section presents the results of our experiments which are based on one month of AFC data from the Société des Transports de l'Outaouais (STO). The fifth and last section summarizes and concludes this paper and discusses future directions.

2 Literature review

We start with the description of AFC system. A state of the art of smart card data enrichment is then presented. Limits to smart card studies are described in the last part.

2.1 Automated Fare Collection data in public transport system

The smart card is used to grant access to the public transit system, according to the faring policy Pelletier et al. (2011). There is some heterogeneity among AFC systems and the accuracy of the data they produce (Erhardt, 2016). Bagchi and White (2004) identify five dimensions attached to smart card data: time information, purchase information, structural information (transportation mode, line, direction), spatial information (not systematically recorded), and the rider's personal information (recorded but never displayed due to privacy concerns).

2.2 Smart card data enrichment

In some faring systems, only a tap-in when boarding is required. Enriching smart card data with alighting locations is one of the main goals in smart card data enrichment. Trépanier et al. (2007) propose a rule-based methodology which relies on a comprehensive description of the trip-chain structure. Figure 1 presents a summary of this methodology: someone boarded at boarding stop 1, the vanishing route is the sequence of stops where the traveler can alight (potential alightings); if his or her next boarding (boarding stop 2) is close to the vanishing road, then it is considered that he or she alights at the closest station of this next boarding (closest alighting). The vanishing route is considered as close if it is within a walkable distance defined by a distance threshold. This methodology has been applied in Gatineau, Canada (Trépanier et al., 2007), Santiago, Chile (Munizaga and Palma, 2012), Brisbane, Australia (Alsger et al., 2015) and other places. To handle tortuous bus routes, (Munizaga and Palma, 2012) also proposed using travel times instead of a distance threshold. The distance threshold sensitivity was explored in (Alsger et al., 2015; Egu, 2015). The rule based approach produces some unlinked trips. To help reduce the number of unlinked trips, He and Trépanier (2015) proposed a kernel density estimation on the historical data of the smart card. Kieu et al. (2015) proposed a similar methodology using the Density-Based Scanning algorithm to use information from historical data to infer more alightings. Zhao et al. (2007) partially validated the rule based approach by verifying counts at an aggregate level and at the bus route ridership level. Later Munizaga et al. (2014) validated the rule based approach using a sample of 601 random public transit riders. They answered a travel survey and agreed to have their smart cards identified. Their reported mobilities were compared with the mobilities processed

through the data produced by their smart cards. For boardings, 98.9% inferred locations were right (small AVL errors due to the GPS system). 84.2% of alighting locations were successfully estimated. Route choices within the Metro system were correct for 85%. These are interesting results, and the article also provides a list of reasons why the alighting estimation failed. This article is very valuable to the research on smart card data considering that previous research rarely included a validation step because there is no dataset which would allow it.

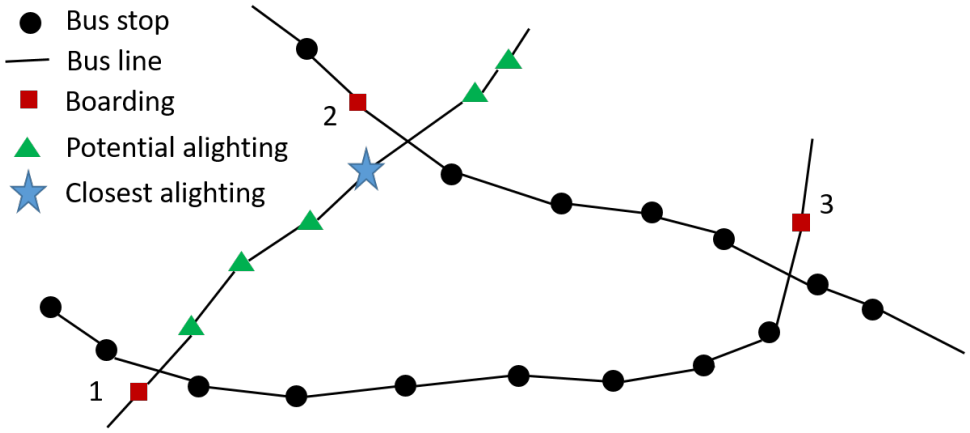

Figure 1: Rational to estimate alighting bus stop for a boarding at bus stop one.

Full knowledge of a trip chain requires knowing boardings, alightings and also activity locations. Figure 2 describes a simple trip chain with two trips (links with diamonds) and three trip legs (arrows). The connection stop (blue box) is only a place to change from bus to subway; it should not be considered as an activity (orange boxes). Knowing activity location allows the production of indicators that are demand oriented similar to those drawn from travel surveys (Trépanier et al., 2009), and to study points of interests and transfer stations (Chapleau et al., 2008). In order to infer activity locations, Chakirov and Erath (2012) proposed and compared two approaches; a rule based approach and a modelling based approach. On the one hand, the rule based approach consists in studying the time interval between an alighting and the next boarding and setting a time threshold; activities lasting six hours to sixteen hours are labeled 'work' purposes and activities lasting one hour to five hours are labeled 'other' purposes. The rule based approach has the main problem of being blind to short activities. It also sets the same time thresholds for all travelers even though traveler behaviours are heterogeneous (Ortega-Tong, 2013). The rule based approach was later reproduced, adapted and validated on other case studies (Devillaine et al., 2012; Munizaga et al., 2014). It was made more complex and detailed by setting new rules about sub-optimality of trips (Nassir et al., 2015). On the other hand, the modelling approach relies on multiple discrete choice models for activity duration, activity start time and destination's land use. It enables a more comprehensive approach, however it was demonstrated using very simple models. The model was calibrated on a dataset including other modes than public transit trips, but no mode choice modelling was done, and it did not include any socio-demographic variables. It enables a great granularity of purposes. The results proved the method to be efficient for the same kind of trip purposes as for the rule based approach. Some papers adopt information fusion approach of AFC data and travel survey data relying on Bayesian methods (Kusakabe and Asakura, 2014; Zhong et al., 2014). They seem provided interesting results, however they are also limited by the spatial accuracy of the travel survey

available.

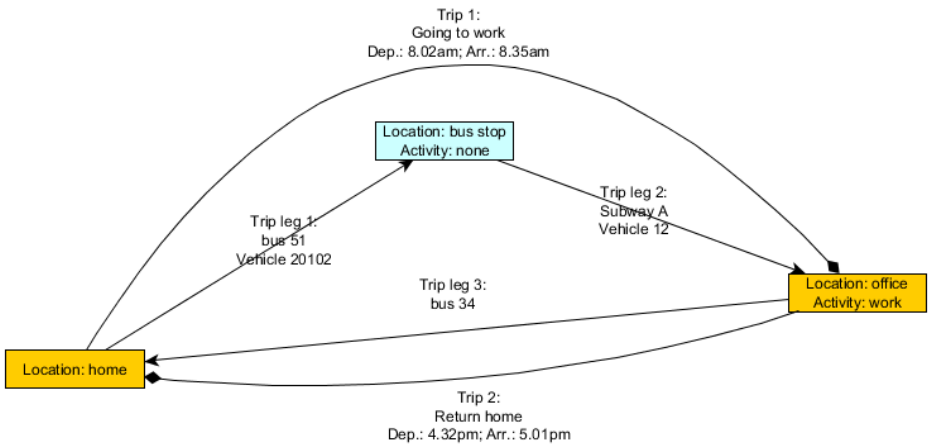

Figure 2: Decomposition of a trip chain.

2.3 Limitations of 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 both spatially and temporally transferable as long as walking distance threshold and transfer time threshold are adjusted to each case study. However, there are some limits to a rule based approach which motivates other strategies. Among the most frequent failures reported by Munizaga et al. (2014) there is overcrowding of buses that delays people from taking the next bus; low service frequency that results in inferring an activity when the traveler was instead waiting for the next bus; short activities impossible to detect with a time threshold etc. These failures are the cause of unlinked trips (when the rule based approach fails to infer an alighting) and wrong destination locations. This creates an inconsistent dataset. Robinson et al. (2014) provide an extensive description of problems that can affect smart card data. For both alighting and activity location inference, the rule based approach is the most documented and developed. It is also an approach both spatially and temporally transferable as long as walking distance threshold and transfer time threshold are adjusted to each case study. However, there are some limits to a rule based approach which motivates other strategies. Among the most frequent failures reported by Munizaga et al. (2014) we find overcrowding of buses that delays people from taking the next bus; low service frequency that results in inferring an activity when the traveler was instead waiting for the next bus; short activities impossible to detect with a time threshold etc. These failures are the cause of unlinked trips (when the rule based approach fails to infer an alighting) and wrong destination locations. This creates an inconsistent dataset. Robinson et al. (2014) provide an extensive description of problems that can affect smart card data.

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, not all transportation systems are equal. Erhardt (2016) presented the case study of Bay area (United States of America) and showed that penetration rates are not equal in various neighbourhoods and in various time windows

(peak hour etc). He emphasizes how the heterogeneity of the penetration rate of the smart card can impact the bus boarding counts estimates. To address this issue, he 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. In transportation systems with no AVL systems, a detailed work has been done by Gaudette et al. (2016) to attach location information using GTFS (General Transit Feed Specification) data and spatial anchor points such as metro stations.

3 Methodology

We present our methodology to infer socio-demographic information in the specific case of smart card data. Our paper contributes to the broader literature around anonymous travel behaviours, which can be derived from GPS information, WiFi connection information, geo-located social networks etc. (Danalet et al., 2014; Poucin et al., 2016). The global theoretical framework we are introducing could be applied to others kinds of anonymous travel behaviour information as well as the smart card. First, we present the available information, then a broad description of the methodology is given, including base hypotheses on which it relies. Finally, 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 is 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 people (according to iTRANS Consulting Inc. (2006)) and it is covering an area of 637 km^2. A very detailed historical description of this organism is available in Blanchette (2009). Our data are from November 2005; there was a bus based service with 56 bus lines (see Figure 3). These are old data, we are interested in developing a generic methodology and the date does not matter. For this period, there are 23,549 smart card holders and a smart card penetration rate of 80% (Blanchette, 2009). The fleet was equipped with an AFC system in 2001 as well as an AVL system. The data available contain, for each smart card transaction:

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

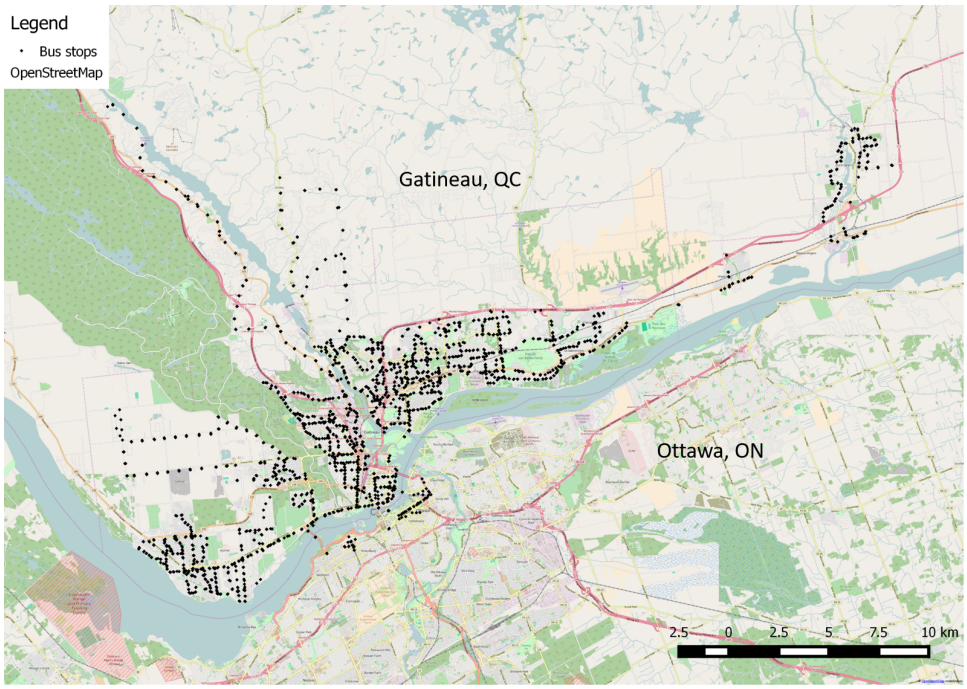

Figure 3: The bus stops of the STO's network. ®OpenStreetMap Contributors

3.1.2 Travel demand survey

The origin-destination survey for the metropolitan area of Ottawa-Gatineau is a land-line based survey with a sampling rate of 5.1%. It is a trip diary survey for the whole day for entire families. The survey was conducted in 2005 between September 21 and November 29. The metropolitan area of Ottawa-Gatineau is divided in two by a river. The STO is running a bus service in the Gatineau area and part of Ottawa downtown. Only 13.5% of all trips (public transit trips and others) in the Gatineau area are made on the STO network, however there are other public transit services. The smart card penetration rate is 80%. Therefore, smart card data are approximately 11% of all trips made in the Gatineau area.

3.1.3 Canadian Census

The Canadian census dataset provides marginal distributions of attributes such as age and gender at the dissemination area level. Dissemination areas (DA) are spatial areas which are designed to be uniform with a population count targeted between 400 and 700 individuals (Canada, 2016). It is the most spatially accurate information, DA's superficy can be as small as 0.016 km^2 in dense urban area. 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 etc. It was held in May, 2006.

3.2 Problem formulation

In order to attach socio-demographic information to the smart cards, we formulate the problem as a maximum weighted bipartite matching problem. The collection of smart cards is one end of the bipartite graph and the collection of people who are potential smart card owners is the second end. The framework and the work flow of the methodology are illustrated in Figure 4. There are five modules:

- population synthesis;
- smart card data processing;
- trip chain choice model;
- creation of the cost matrix;
- distribution of smart cards.

The population synthesis module (module 2) produces the collection of potential smart card owners. The smart card data processing module (module 3) produces the collection of smart cards with enriched information (from data points to trip chain information). Note that module 3 could be traded against any module that takes anonymous travel behaviours and format it up to the level of trip chain information). The trip chain choice module (module 4) calibrates the trip chain model based on the travel survey. In module 1, we use the utility functions of the trip chain behavioral choice model to weight the links between the two previous collections and create the cost matrix which represents the actual matching problem. We solve the matching problem in module 5, using the Hungarian algorithm (Anderson et al., 2014). For each part we relied on existing theory, the novelty is the way we make them work together in a consistent framework and the unique application to smart card dataset enrichment. Each part can be improved independently. We explain the modules in the following parts.

3.3 Module 1: constructing the cost matrix

Our data evolves in a complex space, but we reduce it to four dimensions for explanatory purpose (see Table 1). First, the spatial dimension. Secondly, the temporal dimension. Third, the socio-demographic attributes. Fourth, the mobility choice dimension. The third and the fourth dimensions are not completely independent since the socio-demographic characteristics can have a causal effect on mobility choice. Because of this causality, associating socio-demographic to observed mobility data is not a trivial process.

We came up with three core hypotheses to harness the strengths of our data (see Table 1).

- H1: The first boarding of a user is within the neighbourhood of his or her 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 that the smart card holder is to be found in the local population around the bus stop.
- H2: Individuals have various travel behaviors with respect to their socio-demographic attributes. Therefore, travel habits provide information about the smart card owner.

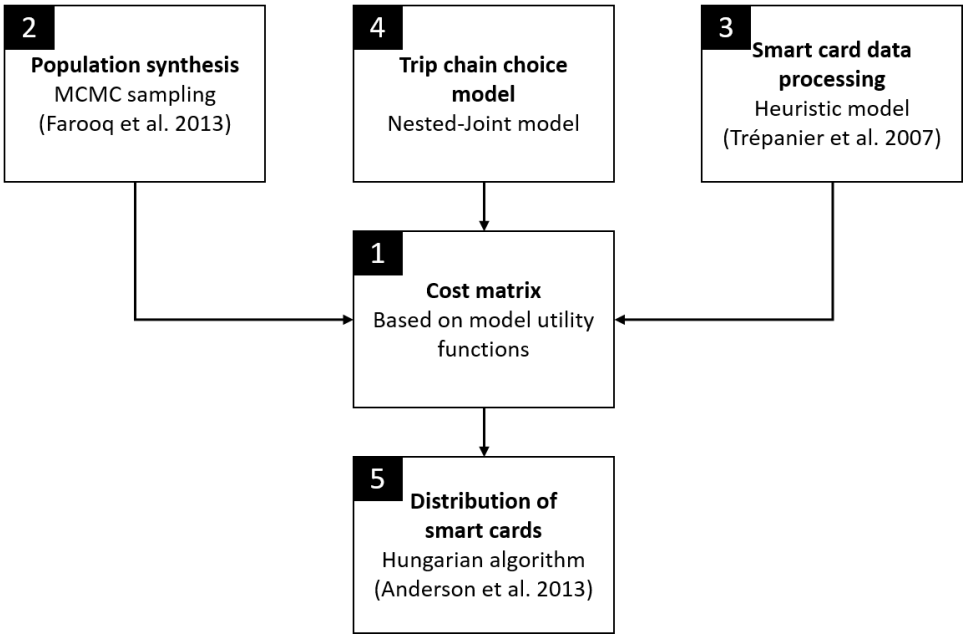

Figure 4: Work flow of the methodology.

Dimension	Spatial	Temporal	Socio-demographic	Mobility choices
Travel Survey	**	**	*****	****
Census	****	*	***	*
PUMS	*	*	*****	*
AFC data	*****	*****	*	*****

Table 1: Problem dimensions. One star represents weak information, five stars represents strong information.

- H3: Public transit riders choose the least expensive fare. This 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, retiree fare etc.).

Figure 5 shows how we are using these hypotheses. From H1, we used the most frequent station for daily first boarding as a living location proxy. It produces localized smart card owners (see Figure 5.a). Thanks to Canadian Census data, PUMS data and travel surveys, the population is known at the dissemination area level. H2 is the core to every behavioral mobility choice model. We calibrate a trip chain choice model and use utility functions to weight links between population and smart cards (see Figure 5.b). A similar methodology was used in Anderson et al. (2014). H3 is used to narrow the choice set in the trip chain choice model: a retiree cannot own a student smart card (in our case study, student fare is for students under 26 years old). We transform our problem into a Maximum Weighted Bipartite Graph problem, with each individual being a potential owner of each smart card. Smart cards are distributed to the population using the Hungarian algorithm. It creates a deterministic and optimized distribution of the smart cards to the population (see Figure 5.c). When the weights for the links are generated out of the trip chain choice model, we use hypotheses

H1, H2 and H3 to reduce the choice set for each agent. It reduces the size of the cost matrix and the time of implementation of the Hungarian algorithm.

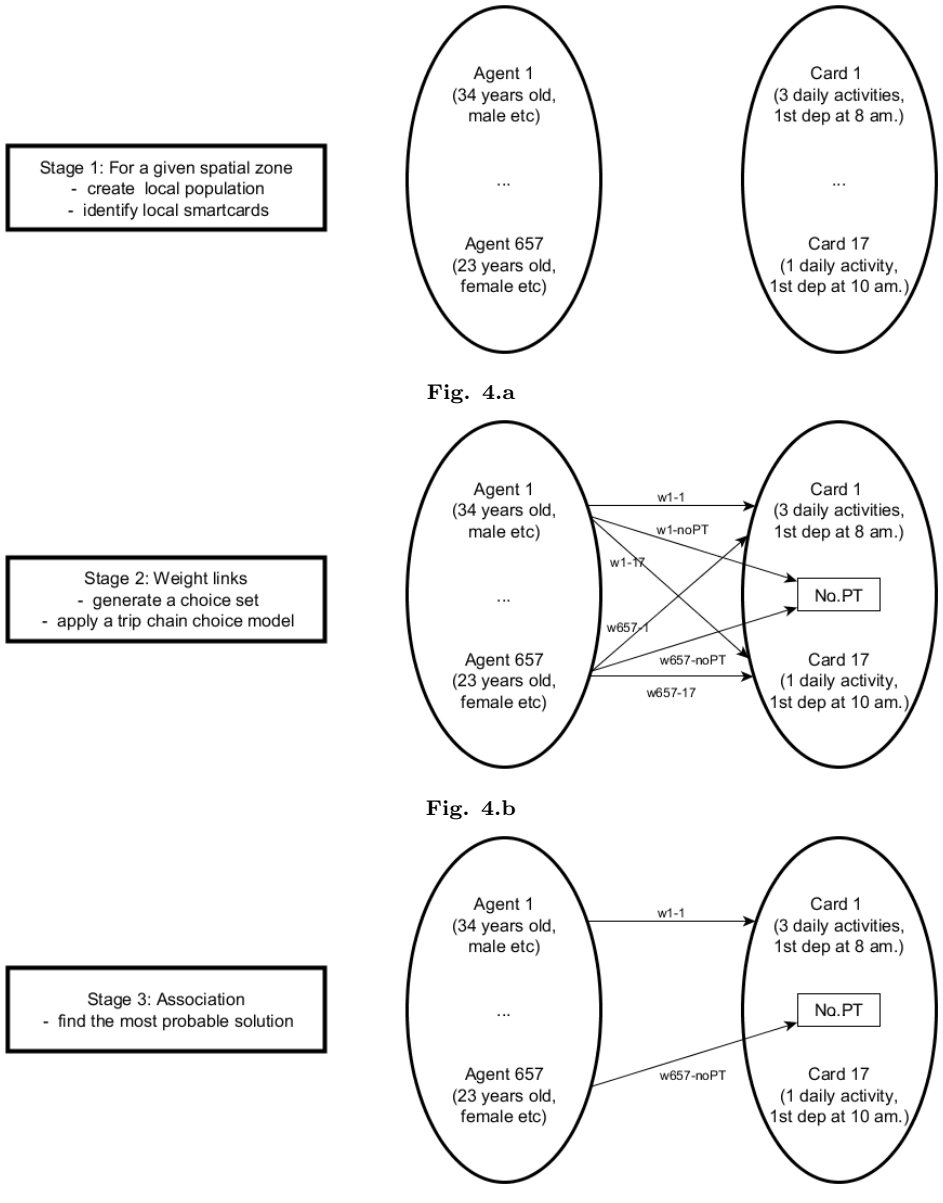

Figure 5: The three stages of the methodological framework.

3.4 Module 2: the 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. We chose to apply a Monte-Carlo Markov Chain sampling method described in (Farooq et al., 2013d,b,a). It relaxes the marginal fitting constraint (traditional constraint in Iteration Proportional Fitting or in Combinatorial Optimization). It proved to be efficient in replicating marginals, and in addition it performs better than traditional techniques to replicate population heterogeneity. We synthesize attributes that will both bring analytic power to the smart card dataset and provide explanatory power for the trip chain choice model. The number of attributes should be limited for two reasons. First, it is causing the computational cost to increase substantially. Secondly, joint distribution of attributes is known through survey data and it may be not enough to provide enough joint information.

Synthesized agent attributes sources can be found in Table 2 (data source is labeled 'X'). When an attribute's distribution is known through different datasets, then we apply the following rules: 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 neighbourhood. Therefore spatial information is the most valuable one. b) If two datasets have the same spatial accuracy, then the dataset that 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.

A better way to approach this would be to take every distributions from the travel survey since it will also be used to calibrate the trip chain model. However, the Gatineau OD survey lacks some basic information (education level, income level, marital status) therefore it needed to be enriched. Also, the age and sex spatial distribution available from Census marginal distributions could be used jointly with the age and sex conditional distribution from PUMS and travel survey using Importance sampling within the Gibbs sampling loop. We did test this, but it did not provide interesting results because the distributions at various spatial aggregation level can be very different.

3.5 Module 3: enriching smart card data with alighting and activity locations

Smart card data are enriched with trip leg's destination using rule based approach (Trépanier et al., 2007). 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) and the hypothesis that trip using public transit are chained. Figure 1 gives a broad approximation of the rational of the methodology: a smart card has boarded at bus stop 1. The potential alightings are the next bus stops along the bus route (in green). Then the smart card was observed boarding at bus stop 2. The closest potential alighting bus stop is in blue. If it is within a walkable distance, then it is labeled as the alighting bus stops.

A time threshold of 30 min between two successive boardings is set to infer an activity location (this time threshold is based on previous studies on the Gatineau area (Devillaine et al., 2012)).

3.6 Module 4: the trip chain choice model

A behavioral choice model has four aspects which must be defined carefully: decision maker (the agent), decision object, decision mechanism and 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

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 cars	$p(car\|age, sex, nPers, occ)$	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

Table 2: Datasets source for attributes distribution (DA: dissemination area (see section 3.1.3), CMA: census metropolitan area).

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. Given our problematic, we have to choose a trip chain definitionf which attributes can be estimated from smart card data. Most smart card data points have 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 (Section 3.5). Ortega-Tong (2013) makes an extensive description of observed trip chains through smart card data and has defined six different discrete and continuous attributes averaged over week-days and week-ends (which makes twelve trip chain attributes): travel frequency, journey start time, activity duration, origin frequency, travel distance, mode choice. On the one hand, the travel habits description can be very detailed. On the other hand, a very detailed description would be a burden for calibrating and applying the trip chain choice model. We have to find a balance between these two aspects. We base our trip chain description on a simplified version of Ortega-Tong (2013). We consider the average time of the first departure and last departure of the day, and we simplified it in three categories: before, during and after peak hour. We consider the average number of daily activities. We also used a specific attribute: the loyalty to STO services. The last attribute is important so we can differentiate STO users for whom we have smart card data from other travelers. It can takes three values: the traveler never used STO services, the traveler is a partial user of STO services (we found missing trips when processing smart card data into trip chain), or the traveler used only STO services to travel (in other words: we didn't find any missing trips in the smart card data). The trip chain description we used can be found in Table 3.

We define public transit loyalty (in our case study: loyalty to STO services) as $\frac{n_{PT}}{n_{PT}+n_{nonPT}}$. 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. The distance threshold (section 3.5) allows us to get an estimator $\tilde{n}_{nonPT}$. If $\frac{n_{PT}}{n_{PT}+n_{nonPT}} \leq 0.95$ we label the smart card as partial public transit user. If $\frac{n_{PT}}{n_{PT}+n_{nonPT}} \geq$ we label the smart card as loyal public transit user.

We have three attributes with three categories and one attribute with four categories (see Table 3). This makes a total of one hundred and height different combinations to describe a trip chain based on smart card data. With our definition of trip chain, there are fifty-four trip chains that either have no activities or don't use public transit services. Since we are using smart card data, there is no point in modelling them. We simplify our model by defining four other choices to handle non public transit riders: car driver, car passenger, active mode, public transit user but non STO rider. This reduces number of available alternatives to fifty-four alternatives and four additional choices other than STO user.

There are many decision mechanisms and most of the applied activity choice models rely on a tree form because they are less demanding when generating the choice set and it is more practical to handle nests than a global architecture. Joint models rely less on causal assumptions. Hess et al. (2012) applied various model structures to the combined choice of fuel consumption and car type. The joint model was found to perform as well as more complex structures (nested and cross nested structures). In our case, we have observed travel patterns (smart card data) and unobserved travel patterns (trip not using the STO network). We intend 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 (see Figure 6). 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 computational efficiency by first segregating the public transit users from

Attributes	Description	Cat.
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 loyalty	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

Table 3: Description of choice attributes

non public transit users and then associating a smart card to a person only on the public transit user population. Usually, activity choice models place mode choice in lower nests. We did the opposite because it was impossible to apply the traditional structure: we are using observed trip chains to build the choice set and we don't have observed choices for non STO nests.

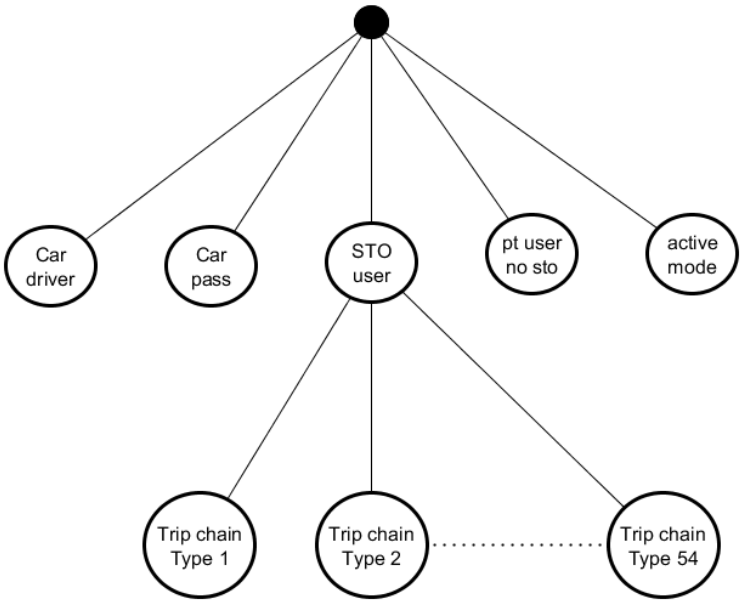

Figure 6: Implemented nested joint model structure.

Not everyone has access to the whole choice set (children and people not owning a car cannot drive). There are two main uses to the trip chain choice model: first, we differentiate STO users from other users (this is the first layer of nest in Figure 6). Second we only need utility functions

to weight the links (see Figure 5.b). 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 ensure that trip chain simulated choices are close to observed choices. There is a difference between an alternative and an observed choice: there may be many observed choices that are labeled as the same alternative, but there are at most fifty-height alternatives. When applying the model to smart cards data we consider all reachable alternatives (within a walkable distance from living location), we also consider the applied fare and whether it is matching with the agent socio-demographic attributes or not, then we assign to each smart card the correspondent alternative utility.

The choice model parameters can include agent specific characteristics (age, gender, income etc.) and alternative specific attributes (land use, accessibility to public transit etc.). Vicinity to public transit is considered when generating the choice set (we are using a walking distance threshold). We tried several level of service indicators for the public transit nest, but it did not improve the model. Final assumptions that we made are based on a comprehensive approach of spatio-temporal constraints of the agents, suggested by the study of iTRANS Consulting Inc. (2006) and by an analysis of the travel survey held in Gatineau in 2005:

- Most of public transit users who didn't use STO are children using school transportation. In our case study, approximately 62% of public transit users who didn't use the STO network are under nineteen years old.
- Mid-aged persons use mainly private mode since they have the physical ability and they can afford to do so. In our case study, approximately 90% of people between nineteen and sixty-four used only car mode for their trip chain during the travel survey.
- Retirees don't use STO services since they are mostly used to car driving, being car passengers, or using some special service that is designed for elders. Approximately 95% of retiree people didn't use STO services during the travel survey.
- Women are more likely to use STO services or to be car passengers. During the travel survey, approximately 30% of women were using STO services (16%) or being car passengers (14%). While only 12 % of men were using STO services and 7% of men were car passengers.
- Couples without children and single persons are more likely to use active mode since they have less time constraints. Approximately, 52% of active trips were made by couples with no child or by singles.
- Elder generation don't use public transit since they are not accustomed to it. For people over sixty-five years old in the travel survey, only 9% of trips were made using public transit.
- Children use STO services: approximately 25% of STO users are under nineteen years old.
- Retirees have fewer time constraint, they are therefore more likely to travel out of peak hours. Approximately 83% of retirees make their first trip of the day after the morning peak hour; and approximately 45% make their last trip of the day (to go back home) before the evening peak hour.
- Home-keepers have the same tendencies: 75% of homemakers make their first trip after the morning peak hour and approximately 43% of them make their last trip (going back home) before the evening peak hour.

id	Dummy	Logic
1	BACK_FROM_SCHOOL	$age \leq 19$ AND $lastDeparture = 0$
2	BIG_FAMILY_1ST_DEP_PEAK	$householdSize \geq 4$ AND $firstDeparture = 1$
60	ELDER_DONT_USE_PT	$age \geq 55$ AND is public transit user
61	HOMEKPR_1ST_DEP_LATE	is home keeper AND $firstDeparture = 1$
62	MID_AGE_USE_PRIV_MODE	age in [25,54]
63	RET_DONT_USE_PT	is retiree AND is public transit user
64	RET_GO_HOME_EARLY	is retiree AND $lastDeparture = 0$
65	RET_LEAVE_HOME_LATE	is retiree AND $firstDeparture = 1$
66	SCHOOL_TRANSPORTATION	$age \leq 19$
67	SMALL_FAM_USE_ACTIVE	$householdSize \leq 2$ AND use active mode
68	WOMEN_ARE_PASS	is woman AND is car passenger
69	WOMEN_USE_PT	is woman AND use public transit
70	WORKER_ARE_NOT_LOYAL	is worker AND use partial public transit
71	YOUNG_USE_STO	$age \leq 19$ AND use public transit

Table 4: Dummy parameters

- Young people leave school earlier than peak hour. Approximately 23% of people under nineteen years old were making their last trip before evening peak hour while only 16% of persons between twenty and sixty-four years old leave before the evening peak hour.
- The bigger a family, the more inter-relational constraints there are, inducing a more constrained schedule, therefore people travel during peak hours. Only 40% of singles are taking their last trip of the day during peak hour. This figure increases with the size of the family; up to 62% for people in families with 5 persons and more.

These assumptions are translated to dummy variables in the utility functions. We are using the following agent specific characteristics: age, gender, household size, occupation. Table 4 shows how the attributes are used to create the dummy variables.

3.7 Modules 5: solving the Maximum Weighted Bipartite Graph problem

Modules 2, 3 and 4 are run in a primary stage. Then modules 1 and 5 are run jointly. Algorithm 1 describes how they are sequenced and how they interact with each other. The implementation relied heavily on the oriented object approach (as described in Trépanier and Chapleau (2001b) and Trépanier and Chapleau (2001a)) to alleviate computational requirement. In this pseudocode, we assume the population synthesis, the model calibration and the smart card data were already processed. In a first step, for each smart card we identify the home location by using the most frequent boarding stop as a living location (see hypothesis H1). We also identify which type of trip chain the smart card relates to among the fifty-height alternatives. We compute several basic statistic data about the trip chain, these statistics will be used later in the application of the choice model. The second step consists in actually giving the smart cards to the population. For each bus station *st*, we extract the local smart cards (smart cards which use *st* as their home bus stop). The population that lives in a dissemination area *zn* which is within a walking distance of the station *st* is considered as the local population. For each agent within the local population, we apply the trip chain choice model using the local smart cards as the available choice set. Then we can construct the cost matrix for the Hungarian algorithm using utility values of the alternatives. We apply the

Hungarian algorithm. The agents who are given a smart card by the Hungarian algorithm are removed from the global population and they won't get a chance to get a second smart card.

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 geography: (s^j), for $j = 1..N_{stops}$
dissemination areas geography:
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 1: Assigning smart cards to the synthetic population.

3.8 Validation

3.8.1 Validation of methodology

We want to validate the socio-demographic dimension of the outcome of our methodology (see Table 1). Bayart et al. (2009) state four levels of validation for information fusion techniques related to information validation:

- preserving marginal distributions
- preserving correlation structures
- preserving joint distribution
- preserving individual values

The best validation possible is at microscopic level (individual values). 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). Spatially there are three different levels of validation possible: macroscopic, mesoscopic and microscopic (individual values). Since ground truth is not available, microscopic validation is not a viable option. Therefore, we decided to work with smart card data leveraged when the travel survey was held. As both datasets are from the same spatio-temporal window we assume it is the best proxy of the ground truth we can get for marginal distributions. We run validations on marginals at macroscopic and mesoscopic scales. We also check how the joint distribution is reproduced by computing the Square Root Mean Squared Errors over tje joint distribution of age x $gender$ x car x occ x $nPers$.

We use three different hypotheses to infer attributes (see Section 3.2). We can develop a sensitivity analysis of these hypotheses by comparing distributions of socio-demographic attributes for:

- OD survey
- smart card holders population while randomly assigned
- smart card holders population while using the local population assumption
- smart card holders population while using the local population assumption and the trip chain choice model
- smart card holders population while using the local population assumption, the trip chain choice model and the fare type

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

3.8.2 Validation of components of the methodology

The population synthesis is controlled using Total Absolute Errors (TAE), Standardized Absolute Errors (SAE) and SRMSE (Anderson et al., 2014; Ballas et al., 1999; Huang and Williamson, 2001).

Destination and activity location inference are difficult to validate because they require to be able to link smart cards to a person *a priori*. Usually they are validated at an aggregate level. 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.

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 presents a quick review of the results and then analyzes the methodology.

4.1 Population synthesis

Population synthesis is operationalized using the open source software SimPSynz (Farooq, 2013; Farooq et al., 2013d; Anderson et al., 2014). We decide not to use Importance sampling to fuse local and global distributions since they can be widely different. We use Gibbs sampling with global distributions and local marginal distributions whenever it is possible (age and gender). Respect of the joint distribution is still controlled by six conditional distributions out of height. For each batch of dissemination areas, and then for each computing thread, the Gibbs sampler was warmed up with a 500 000 draws. Then we sampled an agent every 1000 draws (skip parameter). The method ran in thirty-height minutes, with fifteen batches and seven logical processors of an i7-4710 2.5GHz processor, and height giga-octets of RAM.

SRMSE distributions, SAE maps and TAE maps are not displayed here. Most dissemination areas (DA) have a SAE beneath 0.2 and a TAE beneath 15 for each attribute's distribution (each DA is around 700 people). 90% of DA have low SRMSE. This is satisfactory results for population synthesis.

4.2 Trip chain choice model

Figure 7 describes 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 (this alternative means: person loyal to public transit - first departure hour during peak hour - one activity - going back home earlier than evening peak hour) are students. This means that our description of the trip chain seems to be sufficiently detailed to allow usage segregation.

The trip chain choice model is estimated using the open source software Biogeme (Bierlaire, 2003) and control file for joint model estimation are automatically produced using a software we developed and made available on a github repository (Grapperon, 2016). Given our framework, for the explanatory variables of the utility functions, we can only use agent specific attributes and accessibility indicators at DA level since alternative specific variables are not known for unobserved trip chains.

We use 80% of the data to calibrate the model and we simulate the model on 100% of the data. General statistics and parameter statistics can be found in Table 6 and Table 5. The fifty-four constants from STO nest are removed to alleviate the table, they all have values between -1 and -2.5 and most of them are around -1.5. General statistics give fairly good results with a Rho-square stats of 0.663. The nested structure is validated since the scale factor for STO nest is higher than any parameter. The scale values were estimated on simpler models than we fixed them to help the calibration achieve significant results. Constants' absolute values are quite high showing

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 5: General calibration statistics

that there are still important phenomenons remaining unexplained. No accessibility indicators are used because it appears that a) there is not enough heterogeneity in access to public transit in the travel survey (the population surveyed is mainly around the STO network) and b) Gatineau is a very central city and trip chain are mostly commuting, therefore richness of the public transit offer is not so important to modal choice, as long as the offer fits the commuting demand.

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
56	C_carDriver	1.00	.	.	.
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
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_LOYAL	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	.	.	.

Table 6: Parameters and statistics of trip-chain choice model

Observed choice	Simulated choice				
	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

Table 7: Confusion matrix of the trip chain choice model (in %)

The use of car driving is not well reproduced for the age category 1 (eleven years old to nineteen years old) because it is including people that are not allowed to drive and people that are allowed to drive. We assume that only people over sixteen are allowed to drive, and we compensate this issue by setting that one out of twenty individuals in this category is given the choice of car driving. The model is not able to make a difference between STO services and other public transit services. It results in over estimating public transit other than STO and under estimating STO share. That is why, for age category 1, 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 8 and 9), however it is over estimating STO mode for age categories five and six (over fifty-five years old) and under estimating car passenger in a similar manner. There are important phenomenons remaining unexplained. The bigger nest (car driving) is well reproduced. 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 and STO nests from car passenger and active mode. They are the smallest nests and they can be seen as close alternatives since they are mainly car driving alternatives. The confusion matrix (see Table 7) shows that between 40% and 50% of simulated choices are identical to observed choices (mostly thanks to the car driving mode). Smaller nests are poorly represented. We do not present an analysis of the model for each case of the STO users nest 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.

4.3 Association

We use the Hungarian algorithm implemented in Java from the SympSinz 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 $500m$. The association process runs in approximately height hours for 28 000 smart cards and a population of 300 000 persons. Close to half the stations have no smart cards associated to. The major explanation is that the STO network in 2005 is mainly used for commuting from suburbs to major trip attractors. Therefore, some bus stops are only used for alighting. 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 associated. 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.

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

Travel survey expectations (80%)	Home location inference results
500	248

Table 8: Smart card users in the study area

- random distribution of smart cards to the global population
- random distribution of smart cards to local populations around each bus stops
- 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..

4.4 Marginal validation

Resulting STO users population are drawn in Figure 10. We also draw the observed population from the OD survey scaled down to 80% to consider smart card penetration rate. We do a random distribution to the global population to have a base reference: is our methodology performing better than no methodology at all? 20% of the travel survey data are not present in the smart card data. It can 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 and therefore assuming that the 80% of the OD survey is reliable is a strong assumption that may induce biases. The STO user population is rather well reproduced. However, it shows multiple tendencies: it is over estimating the number of single person who uses STO. 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. On this particular point, smart card information is more reliable than the travel survey. This support the fact that we need 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, excepted for the age category 1 (twenty to twenty-four year old) and for the age category 6 (over sixty-five years old).

Another important marginal check can be done at the mesoscopic level. For one neighborhood, surrounded by nature and highways (see Figure 11), we have around 6,000 inhabitants. Our methodology infers that there are 236 smart card users in this neighborhood. After considering weighted travel survey information, we are expecting approximately 500 public transit users in the same area (see Table 8). The smart card penetration rate is 80%. As a simplistic smart card ownership model, we are assuming that this penetration rate is homogeneous over the population. Therefore we are expecting 400 smart users in the area. The smart card counts don't match (see Table 8). We are using a very simple home location model and a very simple smart card ownership model. Mesoscopic marginal validation is impossible at this point.

4.5 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 the internal consistency of the distribution within the socio-demographic dimension. We can see from the results (see Table 9) that there is a slight improvement especially when using the third hypothesis (fare matching). But the activity choice model and the location hypothesis does not seem to improve the SRMSE measure. This reveals that it is really important to have at

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

Table 9: Square Root Mean Standard Error results for the joint distribution of *age* x *gender* x *occupation* x $numberofcars$ x *householdsize*.

least one socio-demographic attribute available in the smart card data to significantly improve the internal consistency of the results.

5 Conclusion

We proposed a behavioural approach to data fusion of smart card data with socio-demographic information. The problem was formulated as a Maximum Weighted Bipartite matching problem. The methodology incorporates four sub parts: population synthesis, calibration of a trip chain choice model, enrichment of smart card data to trip chain level and association of smart card to the population. Major improvements could be performed on each part. We applied the methodology to the case study of Gatineau (Canada). It allowed us to highlight strengths and weaknesses of the methodology. The results showed that a) the methodology is a step in the right direction since it has a noticeable impact on distribution estimates and b) the methodology requires more work, especially on the trip chain choice model, and more validation, so the results can be used.

There are various sources of errors and biases which may occur. They can be summarized into three categories. (A) Errors related to the data: we are using multiple datasets and each of them include its own biases. Surveys have major limitations that we cited in introduction. AFC system failures result in poor quality smart card data. (B) Our work incorporate four distinct parts from four various theoretical field. Each of these parts brings in errors and biases. None the less, it also means that we have four distinct ways to improve the methodology. (C) We had to make some assumptions that may be too simple (walking distance threshold, faring policy etc) and it induces biases very proper to our methodology. Table 10 described more in detail the various sources of errors.

We summarize here the main opportunities we intend to investigate for future work. The trip chain choice model would benefit from further detailed modeling such as including land use information and time of day. Our heuristic approach to infer destinations is also adding several constraints. A more complete trip chain model including land use and time of day variables would allow to release those constraints by taking inspiration from Chakirov and Erath (2012). They compare a heuristic model with a choice model to infer activity locations. The choice model is performing as well as the heuristic model and we could use a similar approach. The inference of home location is currently based on a very simple heuristic (a distance threshold value). It can be improved further by investigating other strategies. A Voronoi's diagram using the stops as the seeds would provide natural boundaries between stops. Given a finite set of points (the seeds), the segments of the Voronoi diagram are equidistant to the two nearest points (see Du et al. (1999) for more details). The population inside a Voronoi's cell would be the local population. It is more consistent since a traveler is more likely to use only one bus stop per bus route. We used a deterministic approach to solve the bipartite matching problem (the Hungarian algorithm). A probabilistic approach such as described in Farooq et al. (2013c) could be applied to our case study. It would not ensure a

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

Table 10: Sources of errors

maximum weight matching, but the matching would be more consistent with the uncertainty of human behaviours. In addition, the proposed algorithm has a linear complexity which makes it more scalable than the Hungarian algorithm. The methodology described in Section 3 can be applied to other anonymous datasets such as WiFi connection, social network etc. If the data contain the spatio-temporal information and if it is possible to link the data points by anonymous user ID, then module 3 (processing of the smart card data into trip chain information) can be interchanged by any model that can take as input anonymous mobility information and process it into trip chain information. It would be interesting to apply the methodology to other types of datasets to assess the costs of transferring the methodology. For data such as WiFi connection, we can expect better results since it is not related to the mode of transportation: smart card allows us to witness only public transit trips.

The smart card technology may not be homogeneously distributed among the population. Therefore implementing a smart card ownership model would improve our work. It is reasonable to expect better results if using more recent data because the smart card penetration rate will be higher and the travel survey will have a better quality. For example, the 2005 OD survey in Gatineau lacks of descriptive variables such as income. The trip chain choice model would benefit from including land use related indicators. The main originality of our work is to propose a conceptual framework that starts from observed mobility (passively collected) and infer socio-demographic attributes by fusing the mobility data to a survey data. This methodology could be applied to various other kind of passively collected data (GPS, bike sharing data, WiFi connection data etc). Using more passively collected data would allow us to reach a more diverse population (not only STO users).

A comprehensive approach linked to origins and destinations in a transportation context is a valuable asset. It allows an accessibility analysis of the public transit system by cohort of population,

by location, by line etc. From a marketing point of view, it allows taxi companies and transport operators to adapt their offer to the local population or to try to extend their market share by targeting specific and localized population. From au urban planning point of view, it enables decision makers and transport planners to evaluate and analyze the impacts of a change made on the network. For example, after building a new metro line, our methodology could monitor live the evolution of usage and population of the new line instead of waiting for the next travel survey. There is a growing trend of last mile transit (microtransit), especially using autonomous vehicles, ridesharing, and even partnering with taxi companies. This is particularly interesting in increasing accessibility in suburbia where conventional public transit is not sustainable. Our methodology can provide key data source for analyzing these new forms of transportation.

References

A.A. Alsger, M. Mesbah, L. Ferreira, H. Safi, Public Transport Origin-destination Estimation Using Smart Card Fare Data, in *Transportation Research Board 94th Annual Meeting*, 2015

P. Anderson, B. Farooq, D. Efthymiou, M. Bierlaire, Associations generation in synthetic population for transportation applications: Graph-theoretic solution. Transportation Research Record: Journal of the Transportation Research Board, 38–50 (2014)

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

K.W. Axhausen, A. Zimmermann, S. Schönfelder, G. Rindsfüser, T. Haupt, Observing the rhythms of daily life: A six-week travel diary. Transportation **29**(2), 95–124 (2002)

M. Bagchi, P. White, What role for smart-card data from bus systems? Municipal Engineer **157**(1), 39–46 (2004)

M. Bagchi, P. White, The potential of public transport smart card data. Transport Policy **12**(5), 464–474 (2005)

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

C. Bayart, P. Bonnel, How to combine survey media (web, telephone, face-to-face): Lyon and rhône-alps case study. Transportation Research Procedia **11**, 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, 587–611 (2009)

M. Bierlaire, BIOGEME: a free package for the estimation of discrete choice models, in *Swiss Transport Research Conference*, 2003

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

S. Canada, *Canadian Census - Fares and payments*, `https://www12.statcan.gc.ca`, 2016. Accessed: 2016-02-10

A. Chakirov, A. Erath, Activity identification and primary location modelling based on smart card payment data for public transport (2012)

R. Chapleau, K.K.A. Chu, Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach, in *11th World Conference on Transport Research*, 2007

R. Chapleau, M. Trépanier, 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

C. Cottrill, Approaches to privacy preservation in intelligent transportation systems and vehicle-infrastructure integration initiative. Transportation Research Record: Journal of the Transportation Research Board, 9–15 (2009)

A. Danalet, B. Farooq, M. Bierlaire, A bayesian approach to detect pedestrian destination-sequences from wifi signatures. Transportation Research Part C: Emerging Technologies **44**, 146–170 (2014)

J. de Dios Ortúzar, L.G. Willumsen, *Modelling transport* (John Wiley & Sons, New-York, USA, 2011)

F. Devillaine, M. Munizaga, M. Trépanier, Detection of activities of public transport users by analyzing smart card data. Transportation Research Record: Journal of the Transportation Research Board, 48–55 (2012)

Q. Du, V. Faber, M. Gunzburger, Centroidal voronoi tessellations: applications and algorithms. SIAM review **41**(4), 637–676 (1999)

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

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, R. Hurtubia, M. Bierlaire, *Simulation based generation of the synthetic populations for Brussels cas study, Chapter 4: Case studies, in SustainCityHandook* (EPFL Press, Lausanne, Switzerland, 2013a)

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

B. Farooq, E.J. Miller, F. Chingcuanco, M. Giroux-Cook, Microsimulation framework for urban price-taker markets. Journal of Transport and Land Use **6**(1), 41–51 (2013c)

B. Farooq, M. Bierlaire, R. Hurtubia, G. Flötteröd, Simulation based population synthesis. Transportation Research Part B: Methodological **58**, 243–263 (2013d)

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

P. Gaudette, R. Chapleau, T. Spurr, Bus Network Microsimulation with GTFS and Tap-in Only Smart Card Data, in *Transportation Research Board 95th Annual Meeting*, 2016

A. Grapperon, *Bataclan*, `https://github.com/billjee/BataclanSlim`, 2016. Accessed: 2016-04-10

L. He, M. Trépanier, Estimating the destination of unlinked trips in public transportation smart card fare collection systems, in *Transportation Research Board 94th Annual Meeting*, 2015

S. Hess, M. Fowler, T. Adler, A. Bahreinian, A joint model for vehicle type and fuel type choice: evidence from a cross-nested logit study. Transportation **39**(3), 593–625 (2012)

Z. Huang, 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, 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 **58**, 193–207 (2015)

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

M.A. Munizaga, 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 **24**, 9–18 (2012)

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

N. Nassir, M. Hickman, Z.-L. Ma, Activity detection and transfer identification for public transit fare card data. Transportation **42**(4), 683–705 (2015)

M.A. Ortega-Tong, Classification of London's public transport users using smart card data, PhD thesis, Massachusetts Institute of Technology, 2013

M.-P. Pelletier, M. Trépanier, C. Morency, Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies **19**(4), 557–568 (2011)

G. Poucin, B. Farooq, Z. Patterson, Pedestrian Activity Pattern Mining in WiFi-Network Connection Data, in *Transportation Research Board 95th Annual Meeting*, 2016

S. Robinson, B. Narayanan, N. Toh, F. Pereira, Methods for pre-processing smartcard data to improve data quality. Transportation Research Part C: Emerging Technologies **49**, 43–58 (2014)

P.R. Stopher, S.P. Greaves, Household travel surveys: Where are we going? Transportation Research Part A: Policy and Practice **41**(5), 367–381 (2007)

M. Trépanier, R. Chapleau, Analyse orientée-objet et totalement désagrégée des données d'enquêtes ménages origine-destination. Canadian Journal of Civil Engineering **28**(1), 48–58 (2001a)

M. Trépanier, R. Chapleau, Linking transit operational data to road network with a transportation object-oriented gis. URISA Journal **13**(2), 23–30 (2001b)

M. Trépanier, K.M. Habib, C. Morency, Are transit users loyal? revelations from a hazard model based on smart card data. Canadian Journal of Civil Engineering **39**(6), 610–618 (2012)

M. Trépanier, C. Morency, B. Agard, Calculation of transit performance measures using smartcard data. Journal of Public Transportation **12**(1), 5 (2009)

M. Trépanier, N. Tranchant, R. Chapleau, Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems **11**(1), 1–14 (2007)

M.F. Yáñez, P. Mansilla, J. de Dios Ortúzar, The santiago panel: measuring the effects of implementing transantiago. Transportation **37**(1), 125–149 (2010)

J. Zhao, A. Rahbee, N.H. Wilson, Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Computer-Aided Civil and Infrastructure Engineering **22**(5), 376–387 (2007)

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

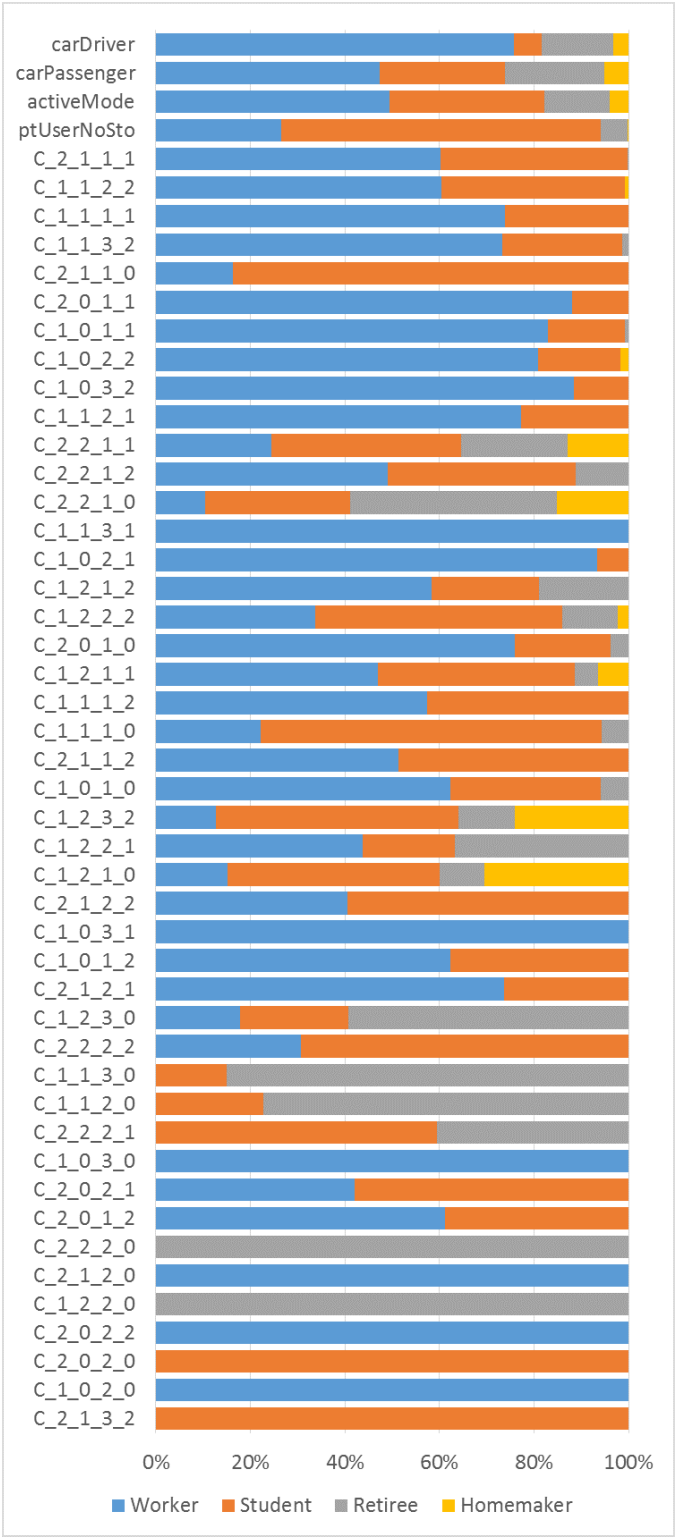

Figure 7: Occupation share for each choice, data: OD survey 2005. The format is C_PTusage_FirstDep_nAct_LastDep. It is also sorted from the most frequent choice (up) to the least frequent choice (bottom).

Figure 8: Marginal distribution of age, gender, occupation, household size and number of cars observed in the OD survey.

Figure 9: Marginal distribution of age, gender, occupation, household size and number of cars simulated for OD survey agents.

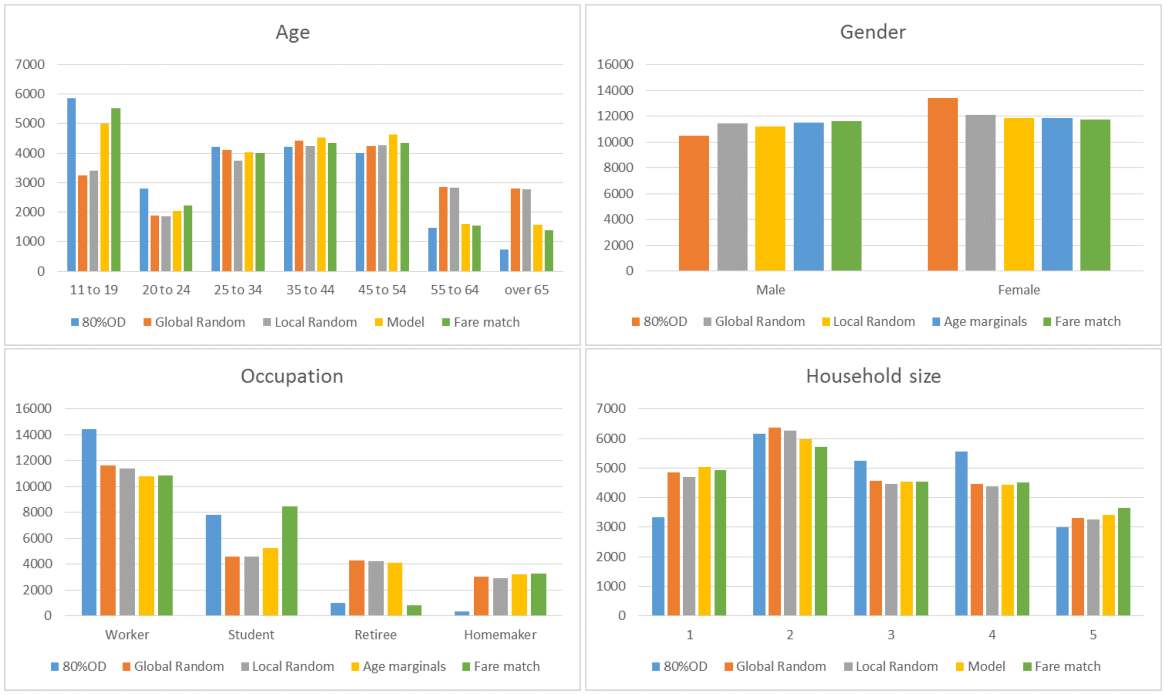

Figure 10: Marginal distributions of age, gender, occupation, household size and number of cars for distributed smart cards.

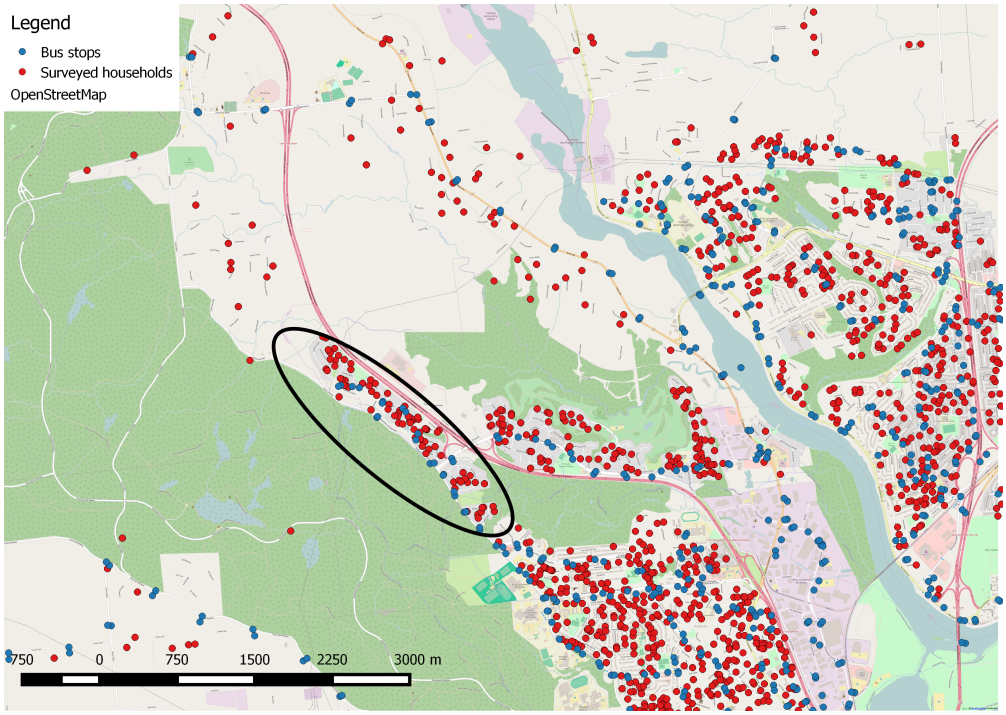

Figure 11: Zone for mezoscopic validation.