DOCTORAATSPROEFSCHRIFT

2011 | Faculteit Bedrijfseconomische Wetenschappen

Creating Synthetic Data Sets for Microsimulation models

Proefschrift voorgelegd tot het behalen van de graad van Doctor in de Verkeerskunde, te verdedigen door:

Juliet NAKAMYA

Promotor: prof. dr. Geert Wets (UHasselt)
Copromotor: dr. Elke Moons (Centraal Bureau voor de Statistiek, Nederland)

universiteit hasselt

DOCTORAATSPROEFSCHRIFT

2011 | Faculteit Bedrijfseconomische Wetenschappen

Creating Synthetic Data Sets for Microsimulation models

Proefschrift voorgelegd tot het behalen van de graad van Doctor in de Verkeerskunde, te verdedigen door:

Juliet NAKAMYA

Promotor: prof. dr. Geert Wets (UHasselt)
Copromotor: dr. Elke Moons (Centraal Bureau voor de Statistiek, Nederland)

D/2011/2451/14

Creating Synthetic Data Sets for Microsimulation Models

to

my dear husband Joshua and daughter Gabriela

‘*Trust in the LORD with all thine heart; and lean not unto thine own*
understanding. In all thy ways acknowledge Him,
and He shall direct thy paths’
Proverbs 3:5-6.

Acknowledgements

Indeed as Ludwig Wittgenstein states, ‘Knowledge’ is in the end based on acknowledgement. Coming to the end of this road has been made possible with the help and support of many people. These few lines intend to express my gratitude to some of them.

First and foremost, I would like to express my appreciation to my promotor Prof. dr. Geert Wets who has guided me throughout this entire research. His thorough guidance and unwavering support is deeply appreciated. I am truly indebted and thankful to my co-promotor Dr. Elke Moons for meticulously supervising my PhD research from its inception to completion. It was a great privilege and an honor for me to have worked closely with her. The sessions we had were very effective, immeasurably contributing to my research aptitude. I express my profound gratitude and thankfulness to the jury members: Prof. dr. Dimitris Karlis, Prof. dr. Koen Vanhoof, Prof. dr. Davy Janssens, Prof. dr. Tom Bellemans and Dr. Mario Cools for their careful reviews, constructive comments and suggestions that have tremendously improved both the content and style of this thesis. I am also thankful for the help of Prof. dr. Ziv Shkedy, who clarified my thinking on different issues in my research.

With candid gratefulness, I extend my appreciation to the whole IMOB staff for their distinguishable support over the four years. They made my working environment a very conducive abode. Special thanks go to Benoît, Konstantinos, Katrien, Els, Enid and Marlies, who supported my research in different ways. I also thank Nadine and Mario, who helped me translate the summary of this book to Dutch and Kristel who diligently helped me with administrative and procedural matters.

With unreserved gratitude I thank Elke Hermans who has been there for me as a friend and sister. Her friendship made my research period so much easier and a lot more interesting, adding more flavor to my wide world. I owe a great depth of appreciation to Papa Luc and Mama Nicole who are my special family in Belgium. Further thanks go to George and Jedidia, Adetayo and Ebun, Karin,

Manuel and Sanneke, Mark and Priscilla and the so many other friends who have been so supportive. To all of those who contributed in one way or another, I can't thank you enough!

I do express my heartiest gratitude to my parents. Daddy and Mummy, thank you for believing in me and supporting me. I hope today, your expectations of me have been somewhat superseded. Thank you for your selfless blessings and sacrifices in all ways. Great thanks go to my dear brothers and sisters who have also greatly stood by me.

This thesis would not have been possible without the help and support of the Almighty God. Without the faith I have in God, my destination would have been uncertain. I thank Him for helping me navigate my way in otherwise difficult circumstances. It was also His love and support that raised me up in my darkest times, when I lost my daughter Favour. It was such a terrible loss I suffered on that very dark day, but the Lord gave me strength to carry on.

I owe profound and earnest thankfulness to my dear husband, which words alone are insufficient to express. It is an honor for me to have had you close by my side throughout this research period. Time and again, I had to dedicate almost all our family time to working on research. For that, obviously, you paid a great price in several ways. Nevertheless, you cheerfully walked with me all the way, lovingly availed your support abundantly and gladly, you always provided a strong shoulder to lean on. When I was feeling low you were there to cheer me up. When the load was heavy you were there to encourage me and you carried my burden when I was fatigued. Your kindness and love never ceases to amaze me. I must say, you are my great inspiration. My dear daughter Gabriela, you are such a great gift. Your sweet smile and laughters always kept me going. For the patience you have exercised towards your full-time working student Mum, I am simply so grateful. At a tender age, one of the first sentences you accurately constructed was ‘Mummy is busy!’. I hope from now on, you can replace this phrase with something like, ‘I am always having a great time with Mummy!’. My sweet daughter, thanks for all your support, I gladly, handover the touch to you, as you start baby class next month!

Juliet Nakamya
Diepenbeek, 28 March 2011

Contents

Acknowledgements **vii**

Contents **ix**

List of Tables **xii**

List of Figures **xv**

1 Introduction **1**

1.1 Introduction 1

1.2 Activity-based Models 2

1.3 Microsimulation Models 3

1.4 Synthetic Data Sets 4

1.5 Organization of Subsequent Chapters 7

2 Motivational studies **11**

2.1 Introduction 11

2.2 The Surveys 11

2.3 The Socio-economic Census 15

2.4 Administrative Sources 16

3 Data integration **19**

3.1 Introduction 19

3.2 Methodology 24

3.3 General guidelines in preparing for data integration 31

3.4 Results and Discussion 32
3.5 Conclusion 43

4 Simulation of travel data 45
4.1 Introduction 45
4.2 Methodology 48
4.3 Trip rates by travel purpose 57
4.4 Trip rates by travel mode 77
4.5 Duration of work trips 82
4.6 Conclusion 84

5 Models for prediction of some travel-related variables 87
5.1 Prediction of car ownership 88
5.2 Car Mileage model 109
5.3 Prediction of personal income and possession of driver's license 116
5.4 Conclusion 119

6 The framework for synthetic populations 123
6.1 Introduction 123
6.2 Data Requirements 128
6.3 The General Procedure for Generation of Synthetic Populations 129
6.4 Measures of Comparison 138
6.5 Discussion and Conclusion 139

7 Creating synthetic populations for Flanders 143
7.1 Introduction 143
7.2 Data 145
7.3 Creating Synthetic Populations through Different Procedures . 146
7.4 The Synthetic Population of Flanders across Years 176
7.5 Conclusion 185

8 Final conclusions 189

A Population synthesis: Implementation 199
A.1 Detailed synthetic population generation model 203

B Data 205

Bibliography **211**

Summary (in Dutch) **229**

List of Tables

2.1 A Comparison of the sample design of the *FHTS* and the *FTUS* surveys 14

3.1 Illustration of the sample data of file A and B for a statistical matching problem 26
3.2 Comparison of the percentage of respondents by socio-demographic factors with respect to the *FHTS*, *FTUS* and the *Synthetic sample* (16-75 Years) 34
3.3 Average number of trips per person per day and the average travel duration per person per day by socio-demographic factors with respect to the *FHTS* and the *Synthetic sample* (16-75 Years) . . . 36
3.4 Percentage of trips by travel goals of respondents (16-75 years) . . 37
3.5 Percentage of trips of respondents by modes of travel 37
3.6 Comparison of descriptive statistics for fusion variables 38
3.7 Correlations between the distance and duration for home to work/school travel variables 42

4.1 Comparison of the subsistence (travel purpose) trip rates between the survey data and the simulated data 63
4.2 Comparison of the subsistence (travel purpose) trip rates between the survey data and the *FIX* simulation procedure data 64
4.3 Comparison of the maintenance (travel purpose) trip rates between the survey data and the simulated data 65
4.4 Comparison of the out-of home leisure (travel purpose) trip rates between the survey data and the simulated data 66

4.5 Modeling the subsistence trip rates using the Poisson regression model with respect to the training, testing and simulated data sets 71
4.6 Modeling the maintenance trip rates using the Poisson regression model with respect to the training, testing and simulated data sets 74
4.7 Modeling the Out-of-home Leisure trip rates using the Poisson regression model with respect to the training, testing and simulated data sets 75
4.8 Modeling the All-purpose trip rates using the Poisson regression model with respect to the training, testing and simulated data sets 76
4.9 Comparison of the Car mode trip rates between the survey data and the simulated data 79
4.10 Comparison of the Bike/Walk mode trip rates between the survey data and the simulated data 80
4.11 Comparison of the 'Other' mode trip rates between the survey data and the simulated data 81
4.12 Comparison of the duration of work trips per participant per day between the survey data and the simulated data 83

5.1 Parameter estimates from the Box-Tiao models for car ownership and vehicle fleet 103
5.2 Criteria for model comparison 103
5.3 Car mileage (in 1000's of km) by categorization scheme across different data sets 115
5.4 Parameter Estimates from the personal income model – Limburg . 118
5.5 Distribution of predicted personal income 118

7.1 Results of the cell counts for the *IPF* iteration routines in estimation of the household-level joint distribution of Flanders 2007 153
7.2 Results of the cell counts for the final *IPF* iteration routine in estimation of the household-level joint distribution of Flanders 2007 154
7.3 Computation performance results for the different synthesis algorithms 155
7.4 Comparison of the performance of different population synthesis Procedures 156
7.5 The distributions of the control variables 160
7.6 The distributions of some household-level non-control variables . . 163

7.7 The distributions of some person-level non-control (unconstrained) variables . . . 165

7.8 Comparing synthetic data based on different sample data sets - distributions of some control variables . . . 167

7.9 Comparing synthetic data based on different sample data sets - distributions of some non-control variables . . . 168

7.10 Household-level joint distribution by multiple runs . . . 171

7.11 Person-level joint distribution by multiple runs . . . 172

7.12 Person-level joint distribution by multiple runs . . . 175

7.13 The population units of Flanders across years . . . 179

7.14 The distributions of the control variables across the synthetic populations . . . 180

7.15 The distributions of some household-level non-control variables . . . 183

7.16 The distributions of some person-level non-control (unconstrained) variables . . . 184

A.1 The steps in implementing the procedure by Guo and Bhat (2007) 201

A.2 The steps in implementing the *IPU* procedure . . . 202

B.1 Household-level joint distribution for the Flemish population in 2001 - $SEE'01$ . . . 205

B.2 Personal-level Joint distribution for the Flemish population in 2001 - $SEE'01$ . . . 206

B.3 Household-level marginal values for the householders aged 18 and above - Flanders 2007 . . . 207

B.4 Person-level marginal values for the householders aged 18 and above - Flanders 2007 . . . 208

B.5 Household-level marginal values for the householders of Flanders 2021 . . . 209

B.6 Person-level marginal values for Flanders in 2021 . . . 210

List of Figures

3.1 Distributions of travel distance for home-school trips for the *FTUS* versus the enriched *enriched FHTS*. 39
3.2 Distributions of travel duration for school-related trips for the *FTUS* versus the enriched *enriched FHTS*. 40
3.3 Distributions of travel distance for home-work trips for the *FTUS* versus the enriched *enriched FHTS*. 41
3.4 Distributions of travel duration for work-related trips for the *FTUS* versus the enriched *enriched FHTS*. 42

4.1 Example of a tree structure. 50
4.2 Regression tree segmentation results for the subsistence travel purpose trips. 58
4.3 *CART* segmentation results for the maintenance travel purpose trips. 60
4.4 *CART* segmentation results for the out-of-home leisure travel purpose trips. 61

5.1 Car ownership proportions and the corresponding forecast values with confidence intervals using different models 101
5.1 Car ownership proportions and the corresponding forecast values with confidence intervals using different models 102
5.2 Forecast of car ownership in Belgium up to the year 2021. 106
5.3 Vehicle fleet proportions and the corresponding forecast values with confidence intervals using different models 108
5.4 Cross validation plot. 112

6.1 General overview of a population synthesis procedure. 130

7.1 Plot of the Synthetic versus the Target Joint Distributions by Different Household Selection Procedures for Population Synthesis shown at both the Household and Person-level 158

7.2 The population by gender across years. 181

7.3 The population by age groups across years. 182

A.1 Detailed model for generating synthetic populations. 203

1 Introduction

1.1 Introduction

Travel is an integral component of every day life. To meet different needs, humans or goods must cover distances. Travel by use of different means of transport thus permits travelers to meet their needs and provides access to other persons, services and locations. In fact, transport itself is a result of needs that can't be met in situ (Gerike, 2007). Transport plays a vital role and its impact on different sectors such as the environment and the economy is crucial. In as much as efficient planning for the transport sector will highly rely on research on travel behavior, it is well known that the behavior of people is generally complex. It is necessary to establish a fundamental understanding of individual travel behavior and its driving forces and influencing factors. Policy-makers, planners and researchers are aiming at measures that permanently shift people's travel behavior towards a more sustainable mobility. To support planning and policy formulation, transportation models can be used to predict travel behavior so as to make better long-term decisions. Transportation models that are activity-based have set the standard for modeling travel demand. Activity-based models often incorporate microsimulation. The input data required in microsimulation models are in turn often large and unavailable and thus frequently necessitating creation of synthetic data sets.

1.2 Activity-based Models

The activity-based approach to travel behavior analysis emerged in the 1970s in reaction to changes in the transportation policy environment. Since then, significant advances in modeling travel demand have been made. The approach provides a rich holistic framework in which travel is analyzed as daily or multi-day patterns of behavior related to, and derived from differences in lifestyles and activity participation among the population (McNally, 2000). As thus, activity-based travel theory is based on two main notions: Firstly, the major idea behind these models is based on the notion that; travel demand is derived from the activities that individuals and households require or wish to perform. Secondly, people face temporal-spatial constraints, functioning in different locations at different points in time by experiencing the time and cost of movement between the locations (Hägerstrand, 1970). Moreover, humans are generally also constrained to return to a home base for rest and personal maintenance. Travel is therefore viewed in a broader context of activity scheduling in time and space. Therefore, by predicting which activities are performed at particular destinations and times, also trips and their timing and locations are implicitly forecast in activity-based models. Activity-based approaches generally aim at predicting the activities conducted, where, when, for how long, with whom and, if travel is involved, the mode of transport used. By assuming the 'activity', instead of the 'trip', as the basic unit for transportation analysis, and incorporating constraints such as interpersonal dependencies among household members, this activity-based approach is suited to estimate people's travel behavior and support the decision making process in transport policy.

Towards the goal of understanding the total impact of transportation planning actions and fully accounting for their impacts on induced trips, a number of studies have adopted activity-based approaches in analyzing individuals travel behavior. Throughout the evolution of activity-based modeling research, the theory has been continuously refined, tests of specific behavioral hypotheses have been set up, and exploration of methods of modeling important aspects of activity-based travel behavior has been conducted. Several successful implementations of the new models have already demonstrated that the activity-based concept is workable. (Golob and

Golob, 1983; Kitamura, 1988; Ettema, 1996) provide extensive reviews of the literature on activity-based travel theory. An overview and description of the main characteristics of the activity-based modeling approach can be found in (Timmermans et al., 2002; McNally, 2000). Over the past couple of decades, several comprehensive activity-based modeling frameworks have been developed in the United States and different countries in Europe. In Belgium, an activity-based framework for Flanders; *FEATHERS* (Bellemans et al., 2010) has been developed in the same light. *FEATHERS* is an acronym for Forecasting of Evolutionary Activity-Travel of Households and their Environmental RepercussionS.

Microsimulation is increasingly playing a major role in the development of demand modeling practice and also drawing more and more attention from the extensive transportation policy and planning community. Currently, several disaggregate land use and activity-based travel models, which represent decisions and actions of individual persons and households, incorporate microsimulations. In these models, synthetic populations are initially created and the prediction of the outcomes for each unit of the population is made. The results can then be aggregated to guide policy related analysis and decision making.

1.3 Microsimulation Models

Micro-simulation modeling techniques are applied at the fully-disaggregate level of persons and households, which convert activity and travel related choices from fractional-probability model outcomes into a series of 'crisp' decisions among the discrete choices (Davidson et al., 2007). The rationale for development of microsimulation models is that many planners or policy makers are interested in the small-area or local implementation of major regional or national policy changes. As a consequence, microsimulation is generally used at the decision making level, where it is of importance to determine for instance, the impact of policy changes on individual units of interest. Activity-based microsimulation methods employ robust behavioral theory while focusing on individuals and households. Theoretically, the microsimulation concept is based on the individual (or household) unit of behavior just as in any disaggregate model. However, the microsimulation process generates discrete choices of the individual (trip purpose, destination, mode, time of day) rather than an array of probabilities for a population segment associated with each available alternative. The ability to simulate travelers individually allows for consideration of complex linkages across multiple trips. Thus, microsimulation can be viewed as an extension of the conventional demand models providing a natural extension of the disaggregate modeling technique in which simulating trip makers on an individual basis allows for complex linkages across multiple trips, ultimately resulting in a better estimation of real-world travel behavior. According to Vovsha et al. (2002), microsimulation comprises several advantages over conventional probability models. It involves gaining a meaningful insight in the explicit modeling of various decision-making chains and timespace constraints on individual travel that allows for behavioral realism in the demand-modeling procedure. It further involves explicit modeling of the variability of travel demand, rather than that of average values. This implies that there is variability of microsimulation outcomes, which can yield full information about the distributions of the travel demand statistics of interest rather than single deterministic estimates or average values. Microsimulation, among the other advantages, also offers a technical benefit related to computational savings in the calculation and storage of large multidimensional probability arrays. It thus allows virtually unlimited explanatory variables (Davidson et al., 2007).

Self-evidently perhaps, activity-based microsimulation models require significantly large and detailed data that are usually not readily available. In some countries such micro-data are freely available to researchers and planners. However, in most countries, such data are unavailable as their provision would infringe on individual confidentiality laws, or if they are available, they are prohibitively expensive to obtain. This has motivated the need to create such data synthetically.

1.4 Synthetic Data Sets

Several approaches (Mohammadian et al., 2010; Auld et al., 2009; Guo and Bhat, 2007; Ye et al., 2009; Beckman et al., 1996) have been taken in practice in attempting to bridge the gap between the limited available data and the high data requirement needs. The approaches generally utilize the different forms of available data to create further data synthetically. Different considerations frequently come into play in seeking and deciding on which approach to take to generate and provide the data that may be lacking. These may include among others: the nature of data already available, the kind of data that are required to be generated and the purpose for which the data are required. Of all the just mentioned considerations, inevitably, the researcher will highly rely on the already available data. Notably, the available data are in many cases limited. Data needs continue to prevail and this tends perhaps also to be pressing for national statistical agencies as data are not only frequently promptly required, but also in good quality, and great detail. One of the main source of data for transportation planning agencies is travel surveys. Evidently, as observed in different practical scenarios of planning and implementations of travel surveys, several problems are encountered. The surveys demand high effort to plan and implement as well as requiring high costs in terms of time, finances and other resources. The collection of survey data frequently exerts a huge burden on respondents, even for relatively short time periods. Thus respondents are becoming less and less interested in completing or even taking part at all in surveys. This presents another problem of increasing non response in surveys. The problems faced in conducting travel surveys together with the extensive data needs of new disaggregate-based approaches to forecasting travel, warrant the development of techniques for supplementing or possibly replacing existent data collection techniques with synthetic data.

One practical solution that has been conceived by many researchers in attempting to tackle the data scarcity problem is to exploit as much as possible all the information that is already available in different data sources. This has been achieved through a statistical matching approach (D'Orazio et al., 2006) of data already collected. Statistical matching techniques have been used in microsimulation modeling (*e.g.*, Cohen, 1991). The techniques involve combining information from two or more data files to construct one file containing information that is not available in any one file by itself. The input data sources are usually micro data sets that contain information on units such as individuals, households or families. This can be done through a micro approach, that involves integrating data sets at record level by combining records representing a similar or the same real-world entity/unit in different data sets. This is done to construct a complete synthetic file with respect to all variables of interest from the different data sources. The synthetic nature of the resultant file is also useful in overcoming the problem of confidentiality in the public use of micro files. For instance, the U.S. Department of Commerce (Federal Committee on Statistical Methodology, 1980) claims that one of the reasons that motivate to use of statistical matching is the legal restriction on exact matching due to the privacy act. In such cases, one tends to rely on statistical matching even when exact matching is available. A synthetic file also provides useful information in a unique set of data on variables previously not jointly observed. This may be of importance to several agencies and researchers. These synthetic files are also sometimes necessary inputs in other procedures such as microsimulation models. Such complete synthetic data sets may also be preferred due to their ease to analyze as compared to two or more incomplete data sets.

As opposed to generating an entire synthetic file, sometimes it is of interest to generate a single or few attributes of interest. For instance some travel related attributes such as trip rates, trip duration, travel mode e.t.c. may be required. In this case, approaches that permit simulation of such attributes would be functional. Simulating data can be useful in substituting or supplementing Household Travel Surveys (*HTSs*), when an existing *HTS* sample is updated and then the updated data are subsequently used as the basis of the simulation (Stopher and Jones, 2003). This may be ideal in data scarcity circumstances as well as in situations of sudden economic, political or social changes that have influence on travel behavior. The updating may be done in spatial as well as in temporal situations. Owing to past findings (Greaves and Stopher, 2000; Stopher et al., 2003; Pointer et al., 2004; Stopher et al., 2005), simulating *HTS* data, much as it is a relatively novel field of research, entails many prospective benefits. Stopher et al. (2005) report that simulated data form a much lower cost alternative as compared to conducting a full new *HTS*.

There are situations where a complete synthetic population is required to be created, mainly based on a target joint distribution and some sample observed data. There has been growing interest in the generation and use of synthetic populations in microsimulation models over the years. Further interest in synthetic populations has been for data privacy reasons. In this case, statistical agencies consider public release of data sets but are faced with a challenge of guaranteeing the confidentiality of survey respondents and offering sufficiently detailed data for scientific use. Agencies, to keep their end of the bargain, they resort to public release of fully synthetic or partially synthetic data.

In general, a synthetic population can be defined as the creation of a model population that is statistically identical to the real population in the area under study. On the whole, a typical population synthesis procedure aims at generating a micro data set that represents the characteristics of the units of interest. For activity-based travel demand modeling, these units are routinely households and the individuals within these households. Achieving this global task involves two main steps: Initially, a demographic distribution of the units is estimated and next, a matching base sample of these units is drawn from a set of the units that is rich in census information. Therefore, once the population synthesizer has estimated a base (or forecast) year distribution, it generates synthetic population units to match the distribution concerning the demographics of interest. This frequently involves applying a suitable procedure, for example a Monte Carlo procedure, to draw the correct or target number of households of each type from a census sample.

1.5 Organization of Subsequent Chapters

Due to the high cost, low response rate and the long-time needed between data collection and availability of the data, few regions or countries can afford collecting household travel survey data as frequently as needed. In practice, even though some regions are collecting household travel survey data on a regular basis, the average interval between two consecutive household travel surveys may be relatively long. In the meantime, emerging modeling techniques (e.g., microsimulation models) that require much richer data sets are becoming available to planners, therefore, increasing the interest in simulating household travel survey data. Moreover, situations exist, in which it may be useful or even necessary to work with the entire population of actors within the microsimulation, rather than a representative sample. In some applications, unless considerable care is taken, the use of a sample may introduce aggregation bias into the forecast. In situations of larger-scale, more general purpose applications; for instance testing a wide range of policies within a regional planning context, the definition of what constitutes a ‘representative’ sample becomes more ambiguous. A sample which is well suited to one policy test or application may not be suitable for another. This is particularly the case when one requires adequate representation spatially (typically by place of residence and place of work) as well as socio-economically. Population synthesis therefore, may well be the best, and in some cases perhaps the only way to generate the detailed inputs required by disaggregate models.

The main aim of this thesis is to highlight and provide insight on existing methodology in the area of creation of synthetic data sets. Some methods will be examined and assessed, in relation to the more *traditional* approach. In addition, formal formulations of the existing new methods that are focused on in this thesis shall be proposed, to contribute to the developing theory of creation of synthetic data sets. In addition, new or modified techniques shall be proposed. Given that the topic handled here falls in a relatively new field of research, a considerable part of the thesis shall be devoted to the assessment of a selected set of new methods. Moreover, inevitably, given the nature of the research, data comparison will be conducted throughout the thesis.

Briefly, the problem at hand will be handled through the following approaches. Data integration, simulation of synthetic individual-level disaggregate travel data, prediction of car ownership for Belgium, development of a car mileage model and finally, an integrated model will be proposed for generating synthetic populations for Flanders. An overview of each of the chapters within the thesis now follows.

Given that the main focus here is on the creation of synthetic data sets, it is logical to start the thesis by laying out the different data sources that provided the basis for achieving this goal. Chapter 2 provides a detailed discussion of the data sets that are used throughout this thesis. These data arise from surveys that are conducted in Belgium. Further data are also available from a socio-economic census that was conducted in Belgium in the year 2000. The data are thus real-life data sets that had to be cleaned and preprocessed. Additional aggregate level data obtained from several public administrative sources were also used in this research. Following these motivational studies, the general framework of the data integration approach in generating synthetic data entities is the subject of Chapter 3. Statistical matching is then used to combine data from two different surveys, whereby travel survey data are enriched with time use data. This work has also featured in Nakamya et al. (2007,a, 2008, 2010). On the other hand, in Chapter 4, the data simulation approach is presented and applied in the case of simulating travel-related responses (Nakamya et al., 2007b, 2009). In Chapter 5, the focus shifts to prediction of some variables that are important as input in population synthesis models so as to generate more refined synthetic populations. The latter variables are also necessary inputs in *FEATHERS* and are also considered important determinants of travel demand. Chapter 6 ushers in, the methodology for creating synthetic data sets. This work is built upon Nakamya et al. (2010). The Chapter outlines a concise review of the framework for creating synthetic populations providing information on data requirements as well as laying down the procedures that are used for generation of synthetic populations for microsimulation. In Chapter 7, the three algorithms for creating synthetic populations that are presented in the preceding Chapter are applied to the available data and results are compared (Nakamya et al., 2009, 2010). Synthetic populations are generated for Flanders for the target years: 2000, 2007 and 2021 and the results are presented and discussed. In the final Chapter of this thesis, Chapter 8, we present some general conclusions and avenues for future research.

2 Motivational Studies

2.1 Introduction

In this Chapter we will present the data used throughout the thesis. The data available in this research arise from different sources. Data are available from surveys, a socio-economic census conducted in Belgium in the year 2001 as well other public administrative sources.

2.2 The Surveys

This Section provides a description of the surveys that give rise to the data that are used in this thesis. The data arise from three main sources: the Flemish Household Travel Survey (*FHTS*) (Zwerts and Nuyts, 2004), the Flemish Time Use Survey (*FTUS*) (Glorieux et al., 2000) and the *SBO* survey (Cools et al., 2009).

2.2.1 The Flemish Household Travel Surveys

The Flemish Household Travel Survey (*FHTS*), also commonly referred to as the '*Onderzoek Verplaatsingsgedrag (OVG)*' was conducted in Flanders, which is the Flemish (Dutch) speaking region of Belgium. For convenience purposes, we shall use the acronym *FHTS* to refer to this survey throughout this thesis. The survey was carried out in the year 2000 (Zwerts and Nuyts, 2004) among the Flemish citizens who account for about 60% of the Belgian population. The field work took place during a period of 12 months among the Flemish citizens aged 6 years and above. Respondents from a stratified sample of 3,027 households comprising 7,626 persons were requested to fill in an individual questionnaire and also to keep a travel diary for two days. Data was also collected from these households using household questionnaires. In the travel diary, respondents recorded their travel activities, modes of transport, duration, location, company of others when traveling and search for car parking. The individual questionnaire included socio-demographic variables as well as travel-related variables. This survey had a response rate of 32% of the households. Other Flemish Household Travel Surveys, with similar sample designs have also been conducted in Flanders. The first *FHTS* was conducted in 1993 and the latest was recently completed in the year 2007. Only the *FHTSs* of 2000 and 2007 will be used in this thesis. As of 2008, a continuous survey was set up (Petermans et al., 2005) to run at least until the year 2013. In this case, when a distinction is required, they will be referred to as *FHTS'00* and *FHTS'07* respectively.

2.2.2 The Flemish Time Use Survey

The Flemish Time Use Survey (*FTUS*) is was also conducted in Flanders in 1999 by the '*Tempus Omnia Revelat* (TOR)' research group of the Free University of Brussels (Glorieux et al., 2000). The *FTUS* survey (Glorieux et al., 2000), was conducted amongst the Flemish citizens and the fieldwork took place between April 15 and October 30, excluding the period between the 15^{th} of July and the 1^{st} of September in 1999. In this survey, $1,533$ Flemish people between the ages of 16 and 75 were asked to record all their activities in a time use diary for a complete week. There were also questions about subsidiary activities, starting and end times, locations, eventual means of transportation, presence of others, conversation partners during the activity and the motivation to carry out the activity. For the activities, the respondent could make use of a pre-coded list of 154 detailed categories of activities, based on the international time-use study (Szalai, 1972). In addition to the diary registration of *FTUS*, individual questionnaires were also presented to the same sample including socio-demographic variables as well as general indicators on time use and cultural participation. Furthermore, respondents were asked their opinion about different social issues. A 28% response rate of individuals was obtained in this survey.

Table 2.1 offers a comparison of the sample design of the *FHTS* (explained in the preceding Section) and the *FTUS* surveys. In Koelet and Glorieux (2007), a description was given on how both the data sets arising from the *FHTS* and the *FTUS* were cleaned and processed, making them comparable to suit the goals of this research. The process involved intensive, in-depth tasks that required a reasonable amount of time to execute. The general goal of this step was to ensure that the two sets of data were in harmony, reconciled and that the data sources were made compatible.

2.2.3 The '*Strategisch Basis Onderzoek (SBO)*' Survey

The '*Strategisch Basis Onderzoek (SBO)*' survey was conducted in Flanders in 2007 during the *SBO* project that investigated 'an activity-based approach for surveying and modeling travel behavior (Cools et al., 2009). The objective of the survey was providing a representative description of the travel behavior of the population in Flanders. The target population in the survey was defined as 'all the people residing in Flanders, regardless of their place of birth, nationality or any other characteristic'. However, the population that was reached by the survey (*i.e.*, the study population), does not cover the target population completely as only private households and no collective households were considered (Cools et al., 2009). This is also similar with the *FHTS* and the *FTUS*.

In the survey, a stratified clustered design was employed. In the design of the sampling scheme, both the coverage of the people in Flanders and the logistic feasibility of the fieldwork were important concerns. In this stratified design, the population was first divided into non-overlapping groups after which in each group a simple random sample was drawn. The clustered part of the design implied that households served as cluster units. The advantage of using a clustered design, was that one did not need to have a full list of individuals at disposal.

Table 2.1: A Comparison of the sample design of the *FHTS* and the *FTUS* surveys

	FHTS	*FTUS*
Research population	Flanders	Flanders (incl. Flemings in Brussels)
Age	6 years and above	16-75 years
Sampling-unit	Households	Individuals
Fieldwork	12 months	+- 5 months
No. of persons	7626	1533
No. of Households	3027	Not applicable
Sampling	Stratified sample (age of household head)	Stratified sample (community)
Contacting procedure	By telephone/post or exclusively by post	Introduction letter and 2 face-to-face visits
Research instruments	Household Questionnaire	
	Individual Questionnaire	Individual Questionnaire
	Travel Questionnaire (2 days/ retrospective)	Diaries (7 days/ simultaneous)

The total number of successful interviews was set at 2500 households in Flanders based on the selected sample size calculation method used. This was a conservative choice made to achieve the desired sample size given the low previously observed response rates in similar surveys. The survey was set up in such a way, that if a household refused to cooperate, there were 4 reserve households that could make up for this household. These reserve households are matched to the first reference household based on the following factors: municipality where the household lives (based on the National Institute of Statistics (*NIS*)-code), gender of reference person, age category of reference person ($<$ 25 years, 25-34 years, 35-44 years, 45-64 years and $\geq$ 65 years) and household composition (number of adults and number of children).

Data were collected from the households using household questionnaires and individual questionnaires. In addition, respondents were asked to fill in a travel diary for seven days. In the travel diary, respondents recorded their travel activities, modes of transport, duration, location, company of others when traveling and search for car parking. The questionnaires inquired about information related to socio-demographic variables as well as travel-related variables. A mixed survey design of using a Personal Digital Assistant (*PDA*) (Kochan et al., 2010) and traditional paper and pencil diaries methods were used to collect detailed information about planned and executed activity-travel behavior of households. The data collection process took place from the year 2005 to 2008.

2.3 The Socio-economic Census

The last socio-economic population census (*SEE'01*) (Belgian Federal Government, 2009b) was conducted in the year 2001 in Belgium. The data arising from this census were utilized in this thesis with the region of interest being Flanders. The data comprise of socio-economic and demographic variables as well as a few travel related ones. The variables of interest in the census consist of both household and person-level characteristics. These include variables such as household size, availability of car(s) in a household, gender, age, occupation status, total hours of work per week, work/school occupation, occupationally active, work status/schedule, flexibility of work, main mode of work/school related travel, school/work location, the number of work/school daily round trips and departure-location for work/school trips. The Flemish data set of the *SEE'01* contains 2,426,614 households with a total of 5,968,074 persons. The census data set is used as a basis for the data integration procedure in Chapter 3 as well as in the procedures for generation of synthetic populations in Chapter 6 and Chapter 7. For purposes of generation of the synthetic population, the census data are used to obtain the household and individual-level multi-way distributions that are used in estimating the joint distribution for the Flemish population for the year 2007. Table B.1 and Table B.2 show the joint distribution of some variables (at the household and individual-level respectively) from the *SEE'01* that are of interest in generating synthetic population. It was observed that the average household size of Flanders in the year 2001 was 2.46. At the household-level, the variables include: availability of car(s) in a household (*HH-AUTO*) which takes on categories yes/no, age of the householder (*HHDER-AGE*) with categories *18-59*, *60plus* and household size (*HHSIZE*) ranging from *1* to *10+*. The aggregation of the age of the householder was dictated by the available data at the time. At the individual-level, gender (*P-GENDER*) and age groups of persons (*P-AGEGRP*) with categories *0-4*, *5-9*,..., *90+* are the variables of interest.

2.4 Administrative Sources

Micro-level data are not always available to the level of detail required. They are frequently not available or accessible for all variables of interest, all the target units of interest and for the required reference years. In such situations, researchers may rely on aggregate-level data to supplement the available data and use them in conjunction to create new data. It is for a similar cause that disaggregate-level and aggregate-level data were sought and used in the different tasks contained in this thesis. As a consequence, marginal data and joint distribution data, at population-level, were obtained from several government sources that are publicly available through the world wide web (Studiedienst Vlaamse Regering, 2009; Planbureau, 2009; ECODATA, Federale overheidsdienst Economie and Energy, 2009; Belgian Federal Government, 2009b). Table B.3 and Table B.4 show household-level marginal values for the householders aged 18 and above and individual-level marginal values for Flanders in the year 2007 respectively (Studiedienst Vlaamse Regering, 2009; Belgian Federal Government, 2009b). These distributions for were required in the procedure for generating a synthetic population for Flanders in 2007 [See Chapter 6 and Chapter 7]. The population consists of over 6 million people within about 2.5 million households. The average household size of this population is observed to be 2.40, indicating smaller Flemish households compared to what was observed in 2001. In general, majority of the households are either single or double households. The population of individuals comprises of more or less equal men and women and the middle aged ($35 - 54$ years), dominate the population. Similar data were also extracted for the population for the year 2021. The distributions are shown in Table B.5 for the household-level and in Table B.6 for the person-level variables.

In some cases, the sought aggregate-level data were not available in Belgium and were obtained from the Netherlands (CBS, 2009). For instance, data were required on car ownership and some related characteristics for the Flemish population over several years. In Belgium, data on car ownership were only available publicly for 2 years; for 1991 and 2001 (Studiedienst Vlaamse Regering, 2009; Belgian Federal Government, 2009b). However, data on car ownership rates in the Netherlands are available from the year 1985 to 2007 (CBS, 2009), which provides a richer data source for our tasks of car ownership prediction in Chapter 5.

Concerning automobile fleet, data are readily available for Belgium from the year 1974 to 2006 (Belgian Federal Government, 2009a). These data were required for modeling vehicle fleet in Chapter 5. A careful exploration of these vehicle fleet data reveals that data from the year 1974 to 1976 could not be relied on as they represented outliers which could not be supported by known facts or economic trends. These data were thus not used in the model building. More data on net taxable income of the population derived from tax forms were also existent from the year 1976 to 2003 (Belgian Federal Government, 2009a) and were used in modeling vehicle fleet.

3 Data Integration

3.1 Introduction

Vital data on travel behavior is often available from different data sources. These sources include: sample surveys, census data records, as well as other administrative data sources. Nowadays, decision making requires as much rich and timely information as possible. For many years, travel surveys have been and still are one of the most important and rich source of the critical information needed for transportation planning and decision making. These surveys are used to collect current data about the demographic, socio-economic, and trip-making characteristics of persons and households and therefore entailing a rich source of data for explaining travel in relation to the choice, location, and scheduling of daily activities. Travel forecasting is then made possible and the ability to forecast changes in daily travel patterns in response to existent social and economic trends as well as new investments in transportation systems and services are further improved.

Data needs are continuously prevalent and perhaps also pressing for national statistical agencies as data are not only frequently promptly required, but also very well needed in good quality, and ample detail. This is especially due to the need for informed decision making and policy formulation. More still, the need for such data and information is increasing over time as the interest to work on several different research objectives also expands. In most cases, the provision of large quality data on travel demand, which is related to the socio-demographic and travel characteristics of individuals and households, largely relies on household travel surveys (HTS). However, HTS are besieged with challenges as pointed out earlier. These constraints make this approach difficult or inappropriate. Household travel surveys are notoriously expensive and require an appreciable amount of time to plan and implement in spite of the current state of increasingly tight budgets. For instance, in Flanders, the data collection cost can be up to 86.37 euros per completed household (Wets, 2005). On addition, travel surveys are faced with non-participation. Researchers are now getting even more concerned about the high response burden imposed on respondents especially due to the fact that response rates are dropping dramatically. In some situations moreover, it is not plausible to obtain the required data by new surveys. Data integration is of major interest as a means of using available information more efficiently. The responder burden encountered in surveys may be substantially reduced. More still, resource requirements in terms of finances and time required in collecting additional data may thus be reduced. At the same time, a considerable amount of resources has already been invested in creation and maintenance of registers, collection of census and various survey data. This is where is becomes useful to identify strategies that are based on amply making use of the already existing data. Data integration, comes in therefore, as a vital tool and goes a long way in providing a better exploitation of these data sources and as well as contributing to the quality of the available data. With the enrichment of a given data source through data integration, transport planning, operations, management and research agencies can be able to explain relationships, predict and estimate parameters of interest more precisely.

Data integration entails combining data residing in different sources and providing users with a unified view of these data. In the integration process, special attention should be mainly focused on maintaining the integrity and reliability of the data. While a significant amount of work has been conducted on data integration (Arellano and Meghir, 1992; Angrist and Krueger, 1992; Winkler, 1995; Lusardi, 1996; D'Orazio et al., 2006) most of the research has been performed outside the transportation research community. Nevertheless, the need to integrate data from different sources in transportation research frequently arises due to several reasons, such as, overlapping data providing different or conflicting information and the aging of sample survey data and thus the consequent need for updating them. Moreover, the need to enrich and enhance the data set with more information in its own right further provides a motivation for data integration in this field. In fact, today, data integration can be used to reduce the required number of respondents or questions in a survey. It also allows for the types of analyses that would be otherwise impossible based on only one input data source. A good example of data integration in practice, is that of the Belgian National Readership survey, where questions regarding media and questions concerning products are collected in two separate groups of 10,000 respondents each, and then fused into a single survey, thereby, reducing costs and the required time for each respondent to complete a survey (Van Der Puttan et al., 2002).

Data integration is a broad field of research that can be viewed from different perspectives. The problem is wide and it incorporates different aspects ranging from technical considerations, data source integration and the actual statistical data integration. A detailed discussion of technical and metadata related integration considerations is given in Denk and Hackl (2003) and will not be discussed here in detail. In Denk and Hackl (2004), a concrete overview of data integration is provided with a special focus on the techniques and evaluation. The study describes different statistical data integration methodologies including exact matching, statistical matching as well as imputation techniques. Statistical matching entails carrying out a statistical integration of information that has already been collected. Statistical matching is also often referred to as 'data ascription', 'data integration', 'data fusion' or 'synthetical matching' in the literature. In European marketing literature and practice, the most commonly used term is 'data fusion' (Rassler, 2002). In this thesis, we mostly use the term 'statistical matching' and occasionally use 'data integration' synonymously. 'Data integration' is also however sometimes used in the broad sense of its direct meaning.

Historically, statistical matching was motivated by the interest in people's consuming behavior especially to improve media targeting (Rassler, 2002). In these studies, running the required large 'single source' panel was often impractical or costs were prohibitively high. Furthermore, a high percentage of non-respondents or poor quality of data was to be expected (OBrien, 1999). Famous statistical offices such as Statistics Canada as well as market research companies especially in Europe have performed or are still conducting statistical matching. An interesting example; a statistical match between the 1970 Canadian Survey of Consumer Finances and the 1970 Family Expenditure Survey, was carried out at Statistics Canada in connection with work on the measurement and comparison of relative distributions of income for several countries (Alter, 1974). Addition of variables was the purpose of this match. Positive experiences with statistical matching have been published over the years in a wide variety of journals or as internal reports or working papers, e.g., Ruggles and Ruggles (1974), Ruggles et al. (1977), Bakker (1990), Roberts (1994), Aluja-Banet and Thio (2001) and Nakamya et al. (2007, 2008). There have been several studies (Arellano and Meghir, 1992; Angrist and Krueger, 1992; Lusardi, 1996) tackling the issue of integration of data from different household surveys in general. Statistical matching is by now a widely used technique in producing empirical studies. The method is used in many observational studies in medical literature (Rosenbaum and Rubin, 1983; Rubin and Thomas, 1992, 1996; Little and Rubin, 2000). Further applications reflected in the field of economics, include, but not limited to, studies by Wolff (2000); Wagner (2001); Keister (2000) and the Urban-Brookings Tax Microsimulation (Rohaly et al., 2005; Kum and Masterson, 2008).

In general, the choice of a data integration technique depends on the goal of the initiative. D'Orazio et al. (2006) identify two main groups of a statistical integration procedure: the macro and micro approaches. Restricting attention to the macro approach, the main interest is on aggregates of the integrated data. The micro approach on the other hand, which is also the approach of interest in this study, involves integrating data sets at record level by combining records representing a similar or the same real-world entity or unit in different data sets. This is carried out to construct a complete synthetic file with respect to all variables of interest from the different data sources. The synthetic nature of the file has been noted to be useful in overcoming the problem of confidentiality in the public use of micro files. The U.S. Department of Commerce (Federal Committee on Statistical Methodology, 1980) argues out the confidentiality motivation, claiming that one of the reasons to use statistical matching is the legal restrictions on exact matching due to the privacy act. In such cases, one tends to rely on a statistical matching even when exact matching is available. Statistical matching can also be used to correct or adjust the distributions of variable values on a data file where there is reason to believe that the distribution on one file is superior to that on the other (Gavin, 1985). Therefore statistical matching can be seen as a method of improving data (Radner, 1981). A synthetic file, furthermore, provides useful information in a unique set of data on variables previously not jointly observed. This could be of importance to several agencies and researchers. These synthetic files are also sometimes necessary inputs in other procedures such as microsimulation models. Such complete synthetic data sets may also be preferred due to their ease to analyze as compared to two or more incomplete data sets.

It goes without saying that, one of the important aspects in data integration is to measure the quality of the match. As D'Orazio et al. (2006) points out, this is not a trivial problem. There is pressing need for research to assess whether statistical matching is an efficient method of extracting information if one is interested in estimation of statistical models. Unfortunately, until now, very few studies (D'Orazio et al., 2006; Rassler, 2002) have investigated the quality of estimates arising from statistically matched files. Furthermore, a matching procedure should achieve some level of validity. Whereas efficiency usually refers to a minimum mean squared error criterion, as is commonly the case in survey sampling theory, validity is mainly focused on data reproducibility and preservation of the original associations and distributions.

The objective of this Chapter is twofold. The first goal of this study is to examine the impact of data integration on some important travel characteristics (Nakamya et al., 2007). In this case, addition of units to provide a larger *Synthetic sample* will be the purpose of this match. Secondly, through statistical matching, another objective is to create a file of micro data with reasonably detailed information on some variables of interest. Consequently, the synthetic file in this case, will entail adding new variables to the sample. The general approach involves combining data from two surveys based on some available important socio-demographic characteristics that are known to have an impact on travel. The quality of the resultant synthetic files will be evaluated by making a comparison between enriched travel data and the original survey data. Regarding the first objective, intuitively, it can be expected that the resultant synthetic sample, will inhibit lower variation as compared to an original sample. This is a reasonable expectation, since the synthetic sample will be larger. Nevertheless, there is no guarantee that this should always be the case. Moreover, one could strongly argue that since the synthetic sample is based on the available data, the variability will be the same. For the second objective, no particular expectations can be claimed with respect to this study. However, optimism can be held that the method will yield satisfactory results as has been shown in some studies in literature (D'Orazio et al., 2006; Rassler, 2002). The data available for these tasks include data from the Flemish Household Travel Survey (*FHTS*) (Zwerts and Nuyts, 2004), the Flemish Time Use Survey (*FTUS*) (Glorieux et al., 2000) and additionally, the Socio-Economic population census (*SEE*) data of 2001 (Belgian Federal Government, 2009b).

In the subsequent Section, the statistical matching methodology is presented and elaborated. A set of general guidelines in preparing for data integration then follows. Results and a discussion are then given. Finally, the Chapter will be wrapped up, with some concluding remarks.

3.2 Methodology

The main components of a micro data integration approach mainly include the exact and statistical matching procedures. Exact matching is used when matching of records belonging to identical entities is the goal. In this case, data sets with substantial overlap, with respect to observed entities as well as variables are integrated. Statistical matching can be used if exact matching is not possible or not essential for the intended usage of the combined data sets (Federal Committee on Statistical Methodology, 1980). In fact, in many practical cases, the data sets have very few or no entities in common and thus, linked records will still refer to similar or synthetic entities (Denk and Hackl, 2004; Rodgers, 1984) as opposed to exact matches. Since exact matching was not feasible in our study, attention will be restricted to statistical matching.

In response to the availability of data currently retrievable from various sources, statistical matching, a relatively new area of research, has been receiving increasing attention. This is of course geared by the existing demand for data. Denk and Hackl (2004) highlighted the methods that comprise statistical matching to include: techniques separating data sets into equivalence classes and then selecting records to be linked randomly; distance measures for the selection of most similar records; and regression-based techniques (Rassler, 2002; Moriarity and Scheuren, 2001; Rodgers, 1984). Statistical matching constitutes a rich variety of applications in areas such as micro-simulations (*e.g.* (Cohen, 1991)), marketing and official statistics. Standard statistical matching framework *e.g.* (Moriarity and Scheuren, 2001, 2003; D'Orazio et al., 2006) usually involve combining information from two or more data files to construct one file containing information that is not available on any one file by itself. It therefore has the practical objective of drawing information piecewise from different independent sample surveys. The input data sources are usually micro data sets that contain information on units such as individuals, households or families.

To put the framework of statistical matching into perspective, let us consider the problem of statistical matching of two independent surveys; A and B. Table 3.1 gives an illustration of the sample data of file A and B for a typical statistical matching problem. The shaded cells represent the unobserved variables in sample A and B, respectively. Assuming that the files from both surveys contain a set of common variables ($\mathbf{X}$). Let file A contain further, a set of variables ($\mathbf{Y}$) not available in B. File B also contains variables ($\mathbf{Z}$) that are not observed in A. File A; the data set that is to be extended is referred to as the recipient file, while B is the donor file. Without loss of generality, let $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ be a random variable with density $f(\mathbf{x}, \mathbf{y}, \mathbf{z})$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, and $\mathcal{F} = \{f\}$ be a suitable family of densities. Thus, let $\mathbf{X} = (X_1, \ldots, X_R)'$, $\mathbf{Y} = (Y_1, \ldots, Y_Q)'$, $\mathbf{Z} = (Z_1, \ldots, Z_S)'$ be vectors of random variables of dimension R, Q and S, respectively.

If we assume that A and B are two samples consisting of n_A and n_B independent and identically distributed observations generated from $f(\mathbf{x{,}y{,}z})$. Then, the first goal of the statitiscal match in this study looks at the problem in the context of adding units to the first sample thus obtaining $n_A + n_B$ units with respect to the $\mathbf{X}$-variables (Nakamya et al., 2007,a). The *FHTS* data serves as the recipient data and the *FTUS* data, the donor file. The resultant data set after this procedure, which can be viewed as a vertical merge, is referred to as the *Synthetic sample* here. The second objective then focuses on extending the data set by adding variables (*i.e.*, the $\mathbf{Z}$-variables) to sample A, the *FHTS* (Nakamya et al., 2010). To avoid confusion, the resultant sample in the latter objective, unless stated otherwise, is referred to as the *enriched FHTS* here after.

Table 3.1: Illustration of the sample data of file A and B for a statistical matching problem

Sample	Y_1	$\dots$	Y_Q	X_1	$\dots$	X_R	Z_1	$\dots$	Z_S
A	Y_{11}	$\dots$	Y_{1Q}	X_{11}	$\dots$	X_{1R}			
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$			
	Y_{n_A1}	$\dots$	Y_{n_AQ}	X_{n_A1}	$\dots$	X_{n_AR}			
B				X_{11}	$\dots$	X_{1R}	Z_{11}	$\dots$	Z_{1S}
				$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
				X_{n_B1}	$\dots$	X_{n_BR}	Z_{11}	$\dots$	Z_{n_BS}

The shaded cells represent the unobserved variables in sample A and B, respectively

Overall, the procedure implemented here closely follows, that described by D'Orazio (D'Orazio et al., 2006). Extensive processing to deal with the discrepancies between the two available data sources has been dealt with in (Koelet and Glorieux, 2007; Nakamya et al., 2007). Notably, before any data integration task can be effected, significant amount of resources are invested in data preparation. Otherwise, if this aspect is not given enough attention, this could jeopardize the integrity of the match. The issues that require attention in preparation for data integration include, but are not limited to: data cleaning; examining the background information on the available data sources; reconciliations of concepts and definitions; re-categorization, re-coding and transformation of variables; and harmonizing time periods of pre-integration data sets. The considerations will be elaborated upon in the next Section. Not surprisingly, perhaps, the data preparation task can be out rightly exhaustive and to a great extent, rather labor intensive. For instance, given a research objective, a data preparation effort may suddenly reveal numerous unforeseen issues even in a data set that has been already officially declared clean. The research by Koelet and Glorieux (2007), was almost exclusively devoted to this task with Nakamya et al. (2007) also further addressing some aspects.

With relevance to the current study, the survey data sets available were weighted with respect to some important variables at the person-level, that were common between the two surveys. The socio-demographic variables that are common between the two data sets include: gender (male, female), age, marital status (married, divorced, widowed and un-married) and education level (primary school, junior high school, high school and college or university). The weights were computed based on the *SEE'01*, which covered population census data. In general, the weights w_k for group k can be calculated as:

$$w_k = \frac{P_k}{p_k} \tag{3.1}$$

where P_k is the proportion of the k^{th} group in the population and p_k is the proportion of the k^{th} group in the sample. An example of a group, for a simple case of two demographic variables; age and gender may be females aged *25-34* years. Due to small observed sample cell values (in some cases, also small expected cell values) that arise when more variables are used, some groups are re-combined and new weights recomputed in this study.

The distributions of variables, for both the *FHTS* and the *Synthetic sample* data are explored by means of descriptive statistics. On the trip level, the data are also weighted for 'day of the week and 'month of the year to investigate whether a sufficient number of people who were able to do some traveling, was questioned in each month and on each day of the week. For the *FHTS* data, only the first day of travel for each respondent is selected since the quality of data for the second day was found to be low with systematic errors (Nuyts and Zwerts, 2001). Clearly, this recommendation by (Nuyts and Zwerts, 2001) practically offers simplicity, but it comes with a price; loss of information, arising from discarding data. However, this does not have much influence on the current research as interest is not in conducting detailed trip-level analyses; for instance, distinguishing in daily travel patterns. Nevertheless, if one was interested in pursuing such objectives, it could be interesting to embark on an intensive data cleaning job, to erode such errors. Furthermore, the researcher could explore some specialized techniques such as those in missing data field. This could possibly still permit use of the the data captured on the second day to investigate such interesting objectives. Drawing attention back to the current study, further decisions were made regarding the data. Regarding the *FTUS* data, to maintain consistency with the *FHTS* data, it was decided to also select one day of time use registration per respondent. This study further focuses on respondents between the ages of 16 and 75 years, since the *FTUS* data comprise of only respondents between this age range. Age was the only weighting variable for which a discrepancy had to be resolved after Koelet and Glorieux (2007)'s work on harmonizing the two data sets. The categorization of gender, marital status and education level were directly compatible.

With the above review and discussion of the data related aspects, we shall now carry on with the description of the statistical matching approach. This approach can be compared to k-nearest neighbor prediction with the donor as a training set and the recipient as a test set. The procedure consists of two main steps. Firstly, given some element from the recipient set, the set of k best matching donor elements is selected. The matching distance is then calculated over some subset of the common or matching ($\mathbf{X}$) variables, in which case, depending on the nature of the variables, probabilistic approaches or standard distance measures such as Euclidian distance can be used. Units of both files can also be grouped into homogeneous subsets according to given common characteristics. These subsets will be referred to as donation classes here. When donation classes are defined, only records in the same group will be considered as possible donors. For example, an individual of a given gender, only records of the same gender will be considered as possible donors when gender is used to define donation classes. This guards against inconsistencies if for the outcome of interest, the values of males and women happen to be different. In general, the donation classes are defined using one or a few categorical variables chosen within the set of common variables $\mathbf{X}$ in the recipient and donor set.

In the matching application carried out in this study, an individual in the recipient file A is matched with an individual in donor file B. A nonparametric micro approach is followed in which a nonparametric imputation strategy is implemented. Generally, the imputation procedures in the hot deck family seem attractive because they do not need any specification of a family of distributions as they are nonparametric and they do not need any estimate of the distribution function or of any of its characteristics. Nevertheless, it is shown that hot deck methods implicitly assume a particular estimate of either a distribution or a conditional mean function (D'Orazio et al., 2006).

In this study, a hot deck method for statistical matching is used to match the records of the two data sources. In this method, each record in the recipient file is matched with the closest record in the donor file, according to a probabilistic measure using the matching variables $\mathbf{X}$. Here, the nearest neighbors were probabilistically selected with more probability to the closer donors. Following D'Orazio et al. (2006), let us take an example of the simplest case of a single continuous variable X_1. In this case, the method ensures that the donor for the i^{th} record in the recipient file A should be chosen such that for a chosen donor record j^*

$$\delta_{ij^*} = \quad \left|x_{1i}^A - x_{1j^*}^B\right| = \quad \min_{1\leq j\leq n_B} \left|x_{1i}^A - x_{1j}^B\right|$$

where $i = 1, 2, \ldots, n_A$, $j = 1, 2, \ldots, n_B$; n_A and n_B are the number of cases in file A and B respectively.

In general, when two or more donor records are equally distant from a recipient record, one of them is randomly chosen. The equation above follows an *unconstrained* matching, as each record in the donor file is allowed to be used as a donor more than once and it is not required that all of the records in the donor file B be used in the match. Limits are usually placed on the number of times a donor record can be used from file B. These limits are imposed to ensure that the (weighted) distributions of the **Z**-variables *brought over* to the recipient file in the match are closely aligned with the distributions in the original file (Ingram et al., 2000). In constrained hot deck matching, each record is allowed to be selected as a donor only once. This therefore, necessitates that the number of donors is greater than or equal to the number of recipients. The main disadvantage of *constrained* statistical matching, however, is due to the nature of rank order matching: some matches may be made over large distances that are unacceptable or undesirable to researchers. Consequently, additional steps must be taken to minimize this problem. Further details together with the advantages and disadvantages of both approaches have been largely discussed in literature (Ingram et al., 2000; Van Der Puttan et al., 2002; D'Orazio et al., 2006)

In several cases, statistical matching relies on the strong assumption of conditional independence. The assumption implies that given the values of variables common to both data sets, variables found only in data set A are independent of variables found in data set B. In other words, we assume that $P(\mathbf{Y}|\mathbf{X})$ is independent of $P(\mathbf{Z}|\mathbf{X})$. A violation of this assumption is difficult to detect in practice because corroborative evidence from alternative data sources generally does not exist. For instance one way to detect a violation could be to use a measure such as the partial correlation $r_{ZY.X}$. However, there is generally no data available on **X**, **Y** and **Z** to compute this (Van Der Puttan et al., 2002).

The variables age, marital status and education level are used as matching (**X**) variables in the statistical matching procedure performed in this study. Gender is used for defining donation classes. These variables are thus available in both files as previously noted. The fusion (**Z**) variables of interest include: the distance for home-school travel (*disthomesch*), duration home-school travel, duration school-home travel, distance for home-work travel, duration home-school travel and duration work-home travel. These are variables that will be added to the *FHTS* (file A). Note that data are integrated at an individual level.

According to D'Orazio et al. (2006), assessing the accuracy of a statistical matching procedure involves several issues including: model assumptions; accuracy of the estimator; representativeness of the synthetic file; and accuracy of estimators applied on the synthetic data. Efficiency is thus also aimed at, when a match is conducted. Regarding validity, after statistical matching, it is the goal to preserve the marginal and joint distribution of the variables in the donor sample, in the statistically matched file. A practical issue, in particular is that, in many cases, as is the case in this study, there exists no third source of data against which to check the validity of the synthetic data set. In such cases, all that is available in terms of quality control is comparison of the distributions of the donated variables in the donor and synthetic data sets. This may be regarded a necessary but insufficient indicator of the quality of the match. However, if the Conditional Independence Assumption is met, we can be confident that the synthetic data set captures the distribution of the donated variables adequately. In this thesis, we work under the assumption that conditional independence holds.

The quality of the statistical match was examined by means of comparisons that are based on scatter plots, histograms, descriptive statistics and correlation coefficients.

3.3 General guidelines in preparing for data integration

In anticipation of a growing reliance on data integration techniques in future, a set of guidelines based on Nakamya et al. (2007); D'Orazio et al. (2006) is laid down here on some considerations when one is intending to perform data integration. In this Section, reference is made to 'data integration' rather than statistical matching, so as to view this in a broad sense. This guidance should be of interest to individuals and practitioners from a wide range of scientific disciplines. The considerations are not only confined to the transportation field, but can also, rather be extended naturally to other areas of research.

- *Examining background information on travel data and other data sources*: Data sources involve sample surveys, census results and administrative sources. An initial step would be to explore the description, specifics about the survey methodologies and the survey design of pre-integration data sets to ensure compatibility of the different data sources.

- *Reconciliations of concepts and definitions*: After identification of the possible data sources to be integrated with travel surveys, the next major step is reconciliations of concepts and definitions of the different sources. The possible sources may be, for example, some closely related surveys that contain socio-demographic information and other travel variables of interest. Matching any two data sources implies a great preliminary effort in terms of time and resources for their homogenization. For example, when dealing with the combination of two surveys, the two sample surveys may refer to the same unit such as household but actually defined differently.

- *Re-categorization, re-coding and transformation of variables*: For the selected data sources, there should be a set of variables in each data set that is similar. However, even then, another source of inconsistence may be due to differences in classification/categorization and definition of variables. In this case, particular care and considerable work should be devoted to the reconciliation of these inconsistencies. In this respect, some variables are re-coded, re-categorized and other variables are substituted by new variables, by transformation of the available information. Variables that cannot be harmonized will not be used in the matching phase.

- *Harmonizing time periods of pre-integration data sets*: Harmony of time frames may directly be ensured by selecting data sets with the same reference period. However, a situation might occur when two sources are characterized by different base periods. Again, in this case, efforts should be made to harmonize these periods. This can be achieved by applying weights so as to update the data to a reference year e.g. by means of *IPF* (Norman, 1999).

3.4 Results and Discussion

The discussion of results is organized into two parts with respect to the objectives of this Chapter. The first Section pertains to enlargement of the *FHTS* sample, where we examine the impact of data integration on some important travel characteristics. It is to be recalled that, for clarity, the resultant data set here is referred to as a *Synthetic sample.* The second Section is focused on enriching the *FHTS* sample with more variables. Consequently, the created synthetic file here, is referred to as the '*enriched FHTS*' sample. The two synthetic files from the two objectives are purposely treated and handled separately, to offer a clear distinction between the corresponding endeavors.

3.4.1 Enlarging the *FHTS* sample

The distributions of some variables for the obtained *Synthetic sample*; a result of a vertical data integration procedure, were compared with the *SEE'01*, *FHTS* and the *FTUS* data. Table 3.2 shows the distributions of some socio-demographic variables; gender, age, marital status and education level for respondents across all the data sets, reflecting percentage counts. The *FHTS* and the *FTUS* data comprise of 6,401 and 1,527 respondents respectively within the ages of 16 and 75 years. For the *FTUS* , only 1,527 respondents out of the original 1,533 survey participants are kept after data cleaning on travel information. Consequently, the *Synthetic sample* contains 7,928 respondents within the age range of 16 to 75 years. The *FHTS* and the *FTUS* data distributions are very close to the population (*SEE'01*) distributions. With weighting, groups that are under-represented in the sample with respect to the population are up-weighted and those that are over represented are down-weighted. The results show that majority of the people are married and the least are widowed. It is also noted that most respondents are aged between 35 and 54 years. This is the biggest part of the active population; the working population that is the backbone of the economy. Regarding education level, most persons attained high school whereas still a few only a certificate of primary education. It is to be noted that a reasonable proportion of the population (32%) is in the age range of 16 to 34 years. The population is are approximately equally distributed with respect to gender. It is easy to see that the distributions of the *Synthetic sample* are similar to those of the population *SEE'01* and as thus similar conclusions are arrived at based on the *Synthetic sample* and the true survey data (*FHTS* and *FTUS*). Bringing in the *SEE'01* provides a form of independent base of comparison with which the results can be validated. This is possible at this stage, since the information required for comparison is available within the *SEE'01*. In other cases however, this may be an unaffordable luxury as the population data are not always detailed in many specialized fields. Only a couple of questions can be covered in some fields that are of great interest to the country, otherwise, it would be impossible to implement the census successifully. The fields covered in the census include health, travel behavior, not to mention the social and economic sectors among others.

We further explore the distributions of some travel-related variables for the *Synthetic sample* data set in comparison to the original travel data set (*FHTS*). Table 3.3 shows the average number of trips per person per day following gender, age group, marital status and education level for the these data. The average travel duration per person per day for the same factors is also shown in the right panel of the table. The *FHTS* and the *FTUS* trip-level data files comprise of 18,125 and 4,178 trips respectively. Thus, the *Synthetic sample* trip-level data file contains 22,303 trips. The results from the combined data for average number of trips per person per day are quite close to those corresponding to the *FHTS* data. Respondents with college or

Table 3.2: Comparison of the percentage of respondents by socio-demographic factors with respect to the *FHTS*, *FTUS* and the *Synthetic sample* (16-75 Years)

Socio-demographic characteristics	*SEE'01*	*FHTS*	*FTUS*	*Synthetic sample*
Gender				
Male	50	49.80	49.63	49.77
Female	50	50.20	50.37	50.23
Age group				
16-34 years	32	31.79	31.77	31.79
35-54 years	39	39.01	39.80	39.17
55-75 years	29	29.20	28.43	29.05
Marital Status				
Married	62	61.42	61.55	61.44
Divorced	7	7.21	7.20	7.21
Widowed	4	4.46	4.27	4.43
Un-married	27	26.91	26.97	26.92
Education level				
Primary school	18	17.84	15.43	17.35
Junior high school	25	25.42	25.96	25.53
High school	33	32.51	34.36	32.88
College or University	24	24.23	24.25	24.23

university degree are noticeably the most mobile group, both with respect to number of trips and duration. Males are also noted to travel more than females and females appear to undertake shorter trips. The results corresponding to duration for the *Synthetic sample* are slightly higher than those for the *FHTS* data. This is possibly due to the fact that in the *FHTS* data, some respondents did not report the corresponding durations for some intermediate trips. For example, if people incorporate a shopping activity on the way home from work, they tend to forget this small intermediate stop when reporting their travel behavior, while in fact this small stop clearly adds some minutes to the duration of their travel. More to this, in *FHTS*, respondents had to fill out the travel diaries retrospectively, making it even more likely to forget some information. It also seems that respondents in the *FTUS* were more cautious in reporting durations of activities since they had to report every activity regarding their time use throughout the diary-registration period. If the government relies only on the data with some under-reported information such as trip durations, it may cease to make infrastructure investment whereas this may be highly necessary. It seems logical therefore, to rely more on the combined data since it corrects for some deficiencies in the original (*FHTS*) data with respect to other available related survey data (*FTUS*).

Table 3.4 shows the percentage number of trips conducted by respondents following travel goals. Here, respondents record their reason for making a travel trip. It is interesting to note that, much as most out-ward trips from home are work related (with business visits included), shopping trips also come in competitively. Moreover, shopping plus leisure (ports/Culture/Relaxation) are by far the main reasons to why people undertake travel activities. A few trips are made due to following education. The results are again observed to be similar between the *FHTS* and the *synthetic sample* data.

What would be of further interest to see, is what modes people use to travel. Table 3.5 shows the percentage number of trips made by respondents following travel modes used. These are also very close for the *FHTS* and combined data. Most of the traveling is done by car, as would be expected, while slow transport (foot and bicycle) comes in second. As can be observed, the share of public transport is rather low in the general population.

If the government needs to make policy decisions regarding to public transport, it needs to be able to rely on accurate data, which is what is aimed at by combining data from different data sources. The previous results are enlightening. It has been observed throughout the results presented above that the synthetic sample provides similar results as those observed in the surveys. This is however, not surprising as the two surveys *FHTS* and *FTUS* were in themselves generally similar. Nevertheless, there were some aspects picked up from each of the surveys. For instance, the *FTUS* is noted to have resulted into more accurate reporting of durations of activities. The *FHTS* is larger and is quite more close in distribution for the socio-demographic data. However, the *FTUS* is not any much worse but combining the two

Table 3.3: Average number of trips per person per day and the average travel duration per person per day by socio-demographic factors with respect to the *FHTS* and the *Synthetic sample* (16-75 Years)

Socio-demographic	Trips		Duration (in minutes)	
characteristics	*FHTS*	*Synthetic sample*	*FHTS*	*Synthetic sample*
Gender				
Male	2.87	2.83	68.40	73.36
Female	2.77	2.78	54.39	59.37
Age group				
16-34 years	3.12	3.13	64.54	69.22
35-54 years	3.13	3.16	67.31	72.55
55-75 years	2.08	1.97	49.71	54.57
Marital Status				
Married	2.87	2.86	61.19	66.28
Divorced	2.77	2.73	61.01	68.76
Widowed	1.95	1.81	45.93	48.72
Un-married	2.88	2.88	64.66	68.97
Education level				
Primary school	1.91	1.84	38.64	43.06
Junior high school	2.54	2.47	54.52	59.75
High school	3.06	3.09	65.47	71.49
College or University	3.66	3.60	82.48	85.57
Overall	2.83	2.81	61.37	66.33

data sets then provides a more solid data set. In general, since the *synthetic sample* relies on more information, the variance, for example for the means and proportions are expected to be lower as compared to using the original travel (*FHTS*) data alone.

Table 3.4: Percentage of trips by travel goals of respondents (16-75 years)

Travel goals	*FHTS*	*Synthetic sample*
Home	37.79	37.03
Work	13.23	13.53
Shopping	13.62	12.35
Business visit	1.89	1.90
Visiting someone	6.54	6.19
Following education	2.06	3.39
Picking/dropping someone	7.03	6.97
Sports/Culture/Relaxation	9.50	9.82
Services (Doctor, bank)	2.72	3.72
Other	5.62	5.09

Table 3.5: Percentage of trips of respondents by modes of travel

Travel modes	*FHTS*	*Synthetic sample*
Foot	10.24	10.31
Bicycle	12.84	12.20
Motorbike	1.77	1.83
Car	64.53	64.54
Public	3.66	3.84
Other/undefined	6.97	7.28

3.4.2 Enriching the *FHTS* sample with more variables (Z variables)

The synthetic file created through statistical matching (*enriched FHTS*) was compared with the original donor file *FTUS* by means of scatter plots, histograms, descriptive statistics, correlation coefficients as well as regression coefficients. Figures 3.1 to 3.4 show the distributions of the fusion variables

(**Z**). These Z variables were variables that were initially not included in the *FHTS* data. The data for these variables are generated within the *enriched FHTS* sample through statistical matching based on the donor file, the *FTUS*. Due to practical limitations, it is not possible to use a donor sample that is larger than the recipient file. Much as this would have been an ideal choice, such data are not available. In Figure 3.1, the distributions of travel distance for home-school trips are shown and those for travel duration for school-related trips follow in Figure 3.2. In Figure 3.3 travel distance for home-work trips are displayed whereas Figure 3.4 finally focuses on travel duration for work-related trips. The histograms show reasonable similarity of distributions between the two sets of data for the respective variables. The descriptive statistics for these variables were also examined. Table 3.6 shows a comparison of descriptive statistics for fusion variables, for which the mean and standard deviations are shown. The mean and interestingly, the standard deviation, are well replicated in the *enriched FHTS*. The next question would be about the correlation between variables. Would the method be able to preserve the correlations that were exhibited within the donor file? In Table 3.7, the correlations between the distance and duration for home-work and home-school travel are displayed. As would be expected perhaps, the original file exhibits a positive significant correlation between distance and travel duration. The correlations revealed in the synthetic file are slightly smaller but also quite pleasingly similar to those in *FTUS* data. Given that the synthetic file is generated without explicitly accounting for these correlations, this result is remarkable.

Table 3.6: Comparison of descriptive statistics for fusion variables

Variable	*FTUS*	*Enriched FHTS*
Distance home-sch	12.74 (19.94)	12.39 (20.21)
Duration home-sch	25.13 (23.27)	24.89 (21.85)
Duration sch-home	25.75 (23.37)	25.65 (21.98)
Distance home-wk	17.46(21.66)	17.39 (20.34)
Duration home-wk	25.18 (21.61)	25.29 (22.30)
Duration wk-home	25.61 (22.59)	25.80 (22.99)

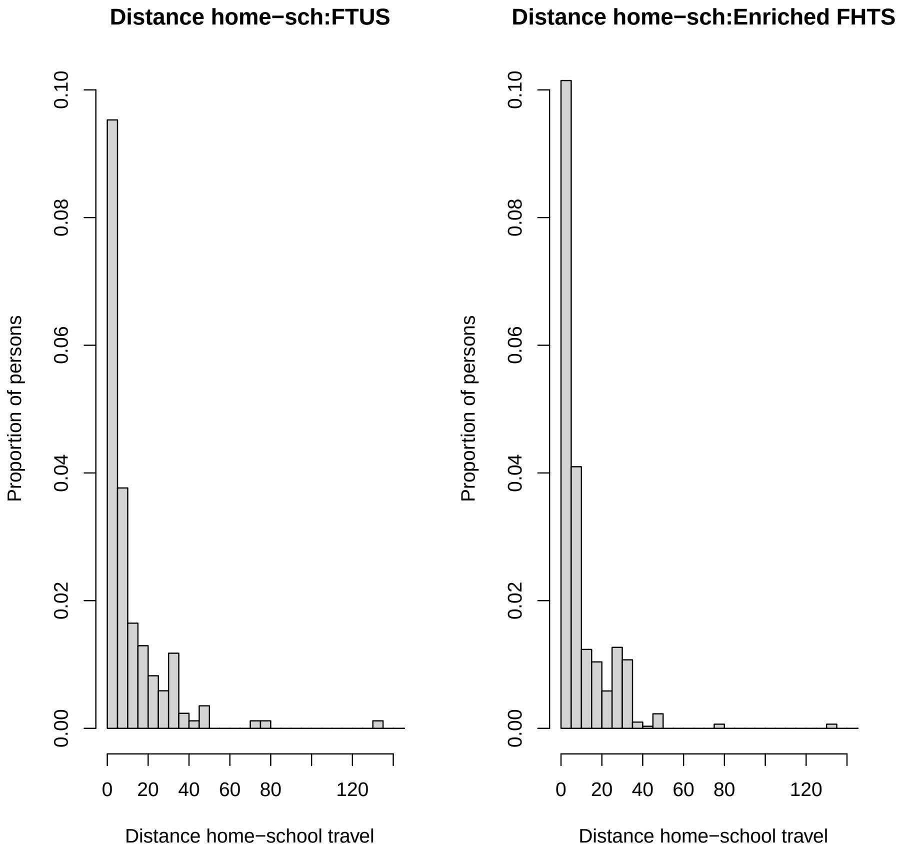


Figure 3.1: Distributions of travel distance for home-school trips for the *FTUS* versus the enriched *enriched FHTS*.

Regarding the argument of validity, the synthetic data created here through statistical matching are quite replicative of actual survey data. This result is right in line with findings in previous research (D'Orazio et al., 2006), in which statistical matching has been investigated. In the current study, distributions and correlations of original survey data have been generally maintained through statistical matching, underscoring the validity of the method. The results thus support statistical matching as a potentially important tool in utilizing data from different sources. It seems promising therefore, to use statistical matching as a tool for integration, to supplement travel survey data. While the approach appears to be elegant, there are still

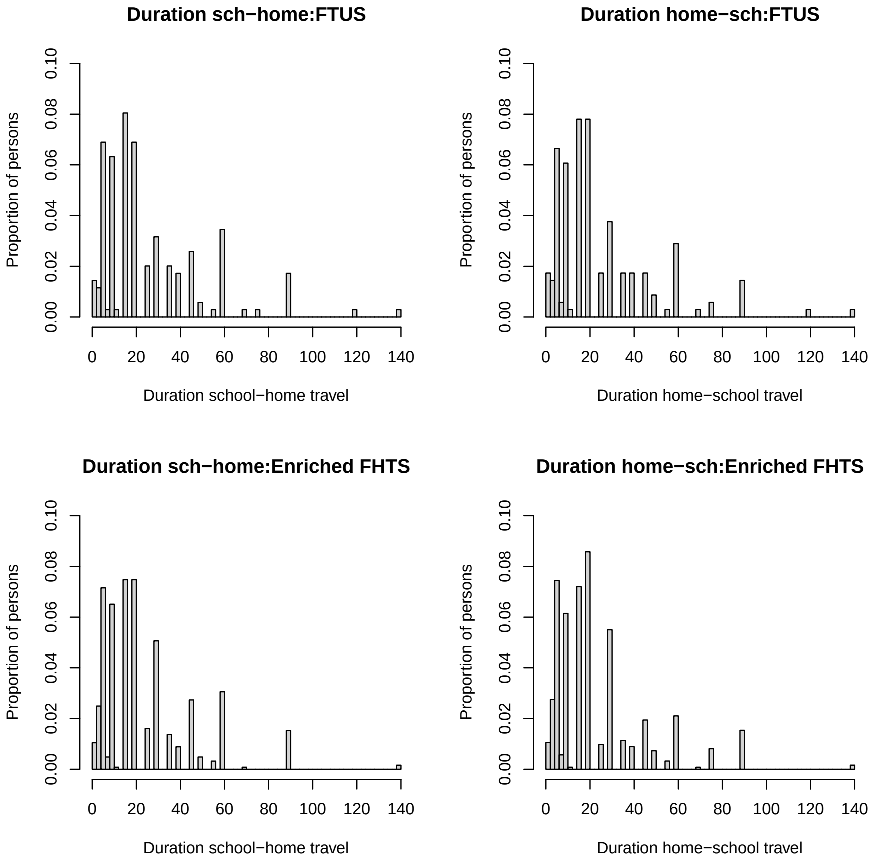


Figure 3.2: Distributions of travel duration for school-related trips for the *FTUS* versus the enriched *enriched FHTS*.

some open problems that can be investigated in future. One important issue is the possible violation of the conditional independence assumption. Roughly speaking, the assumption assumes here that information on the variable X is sufficient to determine Y and Z. It would be interesting to investigate the effect of violations of the conditional independence assumption on the quality of a statistical match. Another problem is of a practical nature. Typically, a donor file would preferably larger than the recipient one. It should be interesting to investigate, when the required data are made available, if using a larger donor file in similar settings would lead to higher qualification of the method. Further research may also follow the line of investigating whether

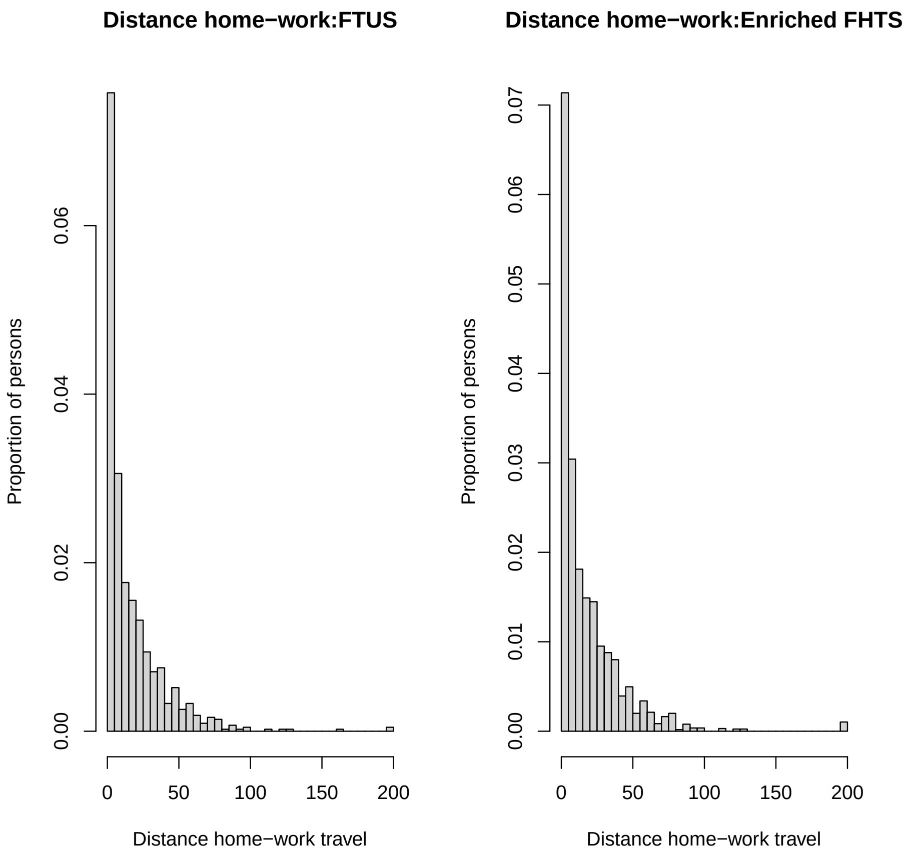


Figure 3.3: Distributions of travel distance for home-work trips for the *FTUS* versus the enriched *enriched FHTS*.

statistical matching is an efficient method of extracting information if one is interested in estimation of statistical models.

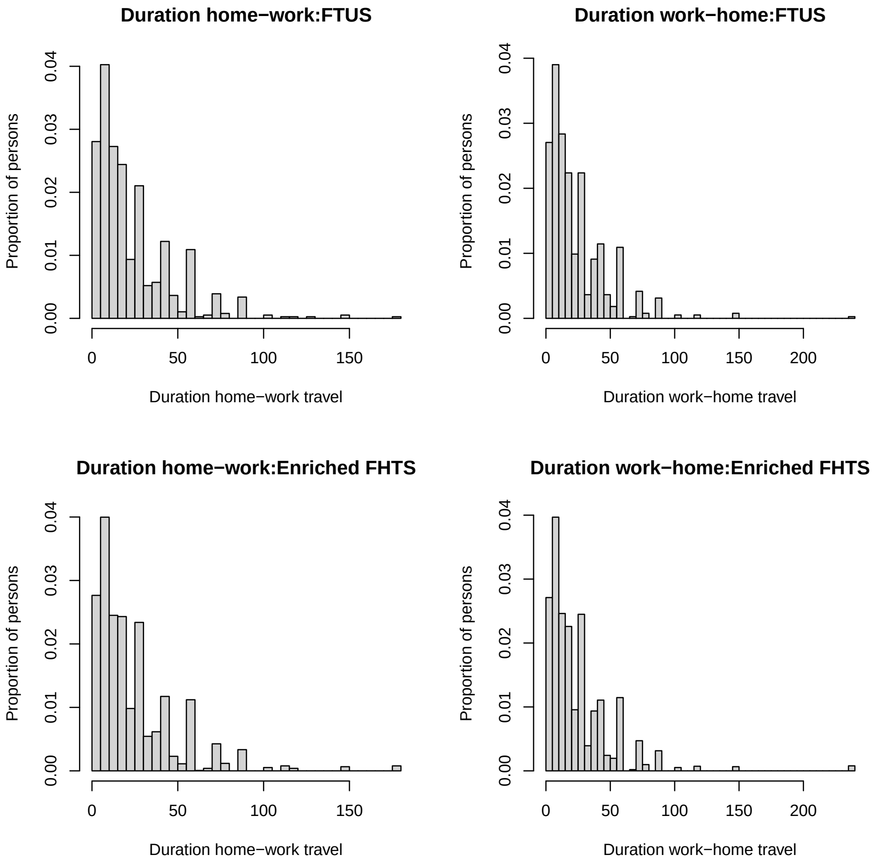


Figure 3.4: Distributions of travel duration for work-related trips for the *FTUS* versus the enriched *enriched FHTS*.

Table 3.7: Correlations between the distance and duration for home to work/school travel variables

		FTUS		*Enriched FHTS*	
Variable		Distance			
		home-sch	home-wk	home-sch	home-wk
Duration	home-sch	0.750		0.726	
	home-wk		0.818		0.809

3.5 Conclusion

Statistical matching constitutes a rich variety of applications in areas including micro-simulations, marketing and official statistics. However, in the field of transportation, there have been limited applications attempts. Furthermore, evaluation of the effectiveness of the technique in the context of estimation has not been adequately investigated so far.

In this Chapter, statistical matching has been conducted enriching the available Flemish household travel survey data with information from a time use survey. The technique has enabled adding new data to existing survey data. The matching has been conducted based on some common variables in both files. The results from comparison of the resultant synthetic file to the original data demonstrated that statistical matching can yield results that are substantially comparable with actual data. Basic statistics such as the mean and standard deviations are considerably preserved. Based on the findings, it seems promising to use statistical matching as a tool for integration, to supplement travel survey data. In the case study taken here, statistical matching has been used to create data relating to travel distance and duration. However, the method can be conceivably used in a broad sense to create other types of variables such as socio and economic variables. This could be potentially useful in providing a richer data set to be utilized in microsimulation models. Moreover, besides the level of detail gained, enlarging a sample through statistical matching further appears to provide a more representative file of the population, which gives more reliable information on the population. The larger sample is valuable in prediction of travel demand and could potentially offer a good base for simulating travel data. This practical experience has also revealed some difficulties relating to data preparation, with respect to which guidelines have been proposed.

It can be arguable that the results obtained here may not be regarded as decisive given that only a single experiment is examined here. It may therefore be interesting to investigate whether the results are generalizable and one way of doing this may be to apply a simulation like the bootstrap. This may be done by generating several bootstrap donor files from the original donor file, and run statistical matching of the same method. It is anticipated that this approach may however involve a heavy computational burden. Future research may also involve integrating information from multiple files using statistical matching while incorporating external information that is not currently available. This will permit a more effective match and validation of the assumption of conditional independence. Moreover, as Rassler (2002) argues, statistical matching may be improved by Bayesian-based techniques, which are quite helpful in overcoming the problem on conditional independence.

4 Simulation of Travel Data

4.1 Introduction

Many of the existing, as well as new disaggregate-based approaches for forecasting travel require extensive data. Also, sometimes, there may be need for information on one or a set of variables, which may not be available in a given survey data set. Moreover, high cost, low response rate and time-consuming data processing, make it prohibitive for many regions to collect household travel survey data as frequently as needed. Therefore, techniques for supplementing the existing data and or data collection procedures have thus turned out to be inevitably vital. Thus far, Statistical matching (Chapter 3) has been discussed as one of the techniques that can be utilized to supplement travel surveys through integrating data from different sources. Simulating data may be another plausible option in attempting to supplement travel surveys. A simulation approach in a slightly restrictive setting, where only one micro-level set of data is available, is attempted here. It is often interesting to investigate if data from a given single source can be replicated. This could be of interest especially in situations where statistical agencies seek provide detailed data without disclosing respondents' sensitive information. Moreover, humans live in a world that is dynamic. When these dynamics turn out to be sudden, there may be serious impact on different sectors of an economy. Sudden economic, political or social changes that may affect travel behavior may mitigate the need to update existing data. This could be performed through data simulation. In such cases, an existing *HTS* sample may be updated and then subsequently used as the basis of simulation. The updating may be effected in spatial as well as in temporal situations.

Simulating *HTS* data, much as it is a relatively new field of research, been considered as an important tool. Researchers have pursued investigation on data simulation in different directions. Interest has been on capturing the complexity and sophistication of human travel behavior as well as improving the quality of data. According to past findings (Greaves and Stopher, 2000; Stopher et al., 2003; Pointer et al., 2004; Stopher et al., 2005; Zhang and Mohammadian, 2008b; Mohammadian et al., 2010), travel data simulation entails many prospective benefits. Stopher et al. (2005) report that simulated data form a much lower cost alternative as compared to conducting a full *HTS*. They argue that the simulated data could be used to generate a database for smaller metropolitan areas in the United States that have insufficient resources to undertake a full *HTS*. More so, due to the high cost, low response rate and time-consuming data processing, few Metropolitan Planning Organizations can afford collecting household travel survey data as frequently as needed (Mohammadian et al., 2010). Simulation enables updating of outdated *HTS* data and further facilitates working with much larger samples. One of the main pitfalls in simulation methods, that is also rather obvious, is that if biased data are used as the basis for simulation, the simulated data will also be biased. In situations of insufficient or lack of *HTS* data, data transferability has also proved to be vital and several studies have been conducted to investigate this. Data transferability is interesting to mention here, since it has been used in close link with data simulation methods. Stopher et al. (2005) utilized transferred data and established that the transfer of distributions from the United States to Australia worked better than expected. The application of Bayesian updating to different transferred travel attributes has been studied (Zhang and Mohammadian, 2008a) and it was shown that the proposed updating approach could significantly improve the quality of the transferred travel data statistics. In investigating the transferability of National Household Travel Survey data, findings (Mohammadian and Zhang, 2007) have demonstrated that some parameters such as trip rates by purpose are easily transferable while others like travel distance can still be transferred with some extra effort involving use of a small local sample and Bayesian updating using Markov Chain Monte Carlo (*MCMC*) simulation. Most recently, Mohammadian et al. (2010) presented a methodology of developing synthetic populations and simulating household travel survey data for areas where actual travel survey data is not available. In the study, Mohammadian et al. (2010) developed transferability models and a Bayesian updating module utilized to facilitate simulating disaggregate household travel data for local areas. A synthetic population for the State of New York (excluding the New York City) was created by a two-stage population synthesis procedure. Then, values of the travel attributes were generated from the updated distributions, by a standard Monte-Carlo simulation. Simulated household travel data for the target area were then created by linking the generated travel estimates to the synthetic population.

The main aim in this Chapter is to simulate synthetic individual-level disaggregate travel data. Thus, an important setting of simulating travel data is considered here. The approach proposed by Stopher et al. (2003) is applied to the case study of Flanders. Building upon the ideas of Stopher et al. (2003), we propose an alternative simulation approach in this study. This approach is also tested on the case study of Flanders and results are compared with those obtained from Stopher et al. (2003) approach. Though considerable research has been doneconducted on simulation methods over the years, to the best of our knowledge, there has not been any systematic research to date that could be used to provide useful directions on the issue of precision from simulation models. In this study, precision of the methods will be evaluated. Contingent on the research goals and the available data in a given study, different travel attributes may be of interest to simulate. Here, interest is particularly focused on the simulation of trip rates and duration data. Three scenarios of data simulation are considered here. Firstly, trip rates data are simulated conditional on people's purpose of travel. In the second instance, we simulate travel trip rates by travel mode and finally, the focus is shifted to simulation of travel duration of work trips. The results of the simulation are then compared with the actual survey data results. It is anticipated that the simulations will results in data that are as good as those from a real survey.

The detailed methodology used in simulating data is presented first, followed by the findings and conclusions.

4.2 Methodology

In this study, simulation procedures are set up to simulate a travel survey data set for a target sample. It is interesting to note that travel behavior comes about as a result of complex interactions and associations among households, individuals, the residential and work areas that they choose and the prevailing geographical and transportation structure at the time. More so, this behavior may change over time. In many simulation studies (Stopher et al., 2003; Janssens et al., 2004; Stopher et al., 2005), the complexity in travel behavior is handled by applying some standard statistical techniques to replicate observed behavior. Such studies have attempted to reproduce the collected *HTS* data and also obtain from models, results similar to those arising from a real survey.

In this study, a procedure similar to that proposed by Stopher et al. (2003) is investigated in to generating simulated travel data (Nakamya et al., 2009). Stopher et al. (2003) procedure for setting up the simulation involves two main steps; categorizing individuals/households and developing distributions from which samples are drawn. These steps are elaborated upon in the next sections. It is important at this point, to mention that the choice of analysis units is an important issue in the procedure. The choice varies mainly, depending on one's research goals and the available data. In travel related simulations, the choice is often between use of individuals or households. Use of homogeneous groups of individuals rather than households has merits, particularly, when dealing with mode choice (Supernak et al., 1983); both theoretical discussion and empirical findings favor the proposed version of the person-category model over household-based models. The reasons given include the fact that it is more practical at the forecasting stage, it has better behavioral background, it requires significantly less data, and it is more compatible with the entire system of individually oriented travel-demand models. Many other researchers (Axhausen and Herz, 1989; Raney and Nagel, 2003) chose to work at the individual level whereas some others (Greaves and Stopher, 2000) opted for household-level in carrying out related research. An argument is put forward that, if the goal of the simulation is to provide a household travel survey data set, which is realistic and plausible, this needs ultimately to be carried out at the level of individual households members (Greaves, 2006). Motivated by these arguments, in this study, we choose to work at the individual level in simulating the number of trips by purpose of travel.

4.2.1 Categorizing individuals

In setting up the simulations the initial step is to categorize individuals into relatively homogeneous socio-demographic groupings with respect to the attribute or outcome of interest, for instance the number of trips made by different individual travel purposes. Classification and Regression Tree (*CART*) models (Breiman et al., 1984) are a flexible tool for estimating the conditional distribution of a univariate outcome given multivariate predictors. The (*CART*) method, an exploratory classification tool is used to categorize individuals into relatively homogeneous groups conditional on their travel purpose. This method has been commonly used in transportation research literature to analyze travel behavior (see *e.g.* Wets et al., 2000; Stopher et al., 2003). Reiter (2005) also suggests use of simulated data and proposes use of Regression trees to generate partially synthetic data. These partially synthetic data are useful when limiting disclosures in releasing public use micro-data. Because of their nonparametric nature, *CART* models have been proposed to impute missing data (Barcena and Tussel, 2000; Piela and Laaksonen, 2001; Conversano and Siciliano, 2002; Reiter, 2005). These proposals primarily use the leaves of trees as imputation classes, assuming the data are missing at random (Rubin, 1976). For example, suppose a single variable Y has data missing at random. A tree is grown using the observed outcomes, Y_{obs}, and all other variables as predictors, then pruned to some desired size. Units with missing Y are placed in appropriate leaves of the tree according to their predictor values, and imputed values of Y are then drawn randomly from the Y_{obs} in the corresponding leaves (Reiter, 2005).

In general, *CART* method is a nonparametric and nonlinear technique, which involves three major steps:

- grow an overly large tree to capture all potentially important splits;
- prune the tree back to the root node to create a hierarchy of sub trees;
- finally, select an optimal-sized tree from this sequence using an independent holdout sample or cross-validation.

Essentially, the *CART* model partitions the predictor space so that subsets of units formed by the partitions have relatively homogeneous outcomes (Chipman and McCulloch, 2000). The method entails making binary recursive partitions of the data (predictors) with respect to the dependent variable of interest to form a tree that consists of different layers of nodes. The series of splits can be effectively represented by a tree structure, with leaves corresponding to the subsets of units. The tree starts from the root node in the first layer, the first parent node. In a binary tree, a parent node is split into two daughter nodes on the next layer. Each of these two daughter nodes becomes in turn parent nodes. This recursive partitioning algorithm continues until a node is terminal and has no offspring. Nodes in deeper layers become more and more homogeneous, less impure, with respect to the response variable.

Let us consider an example for illustration purposes. Consider a tree structure for a univariate outcome Y and two independent variables, *Gender* and *Age*, as presented in Figure 4.1. Binary splits are first performed on *Gender*. Within the 'male' group, units are further split by age. Thus, units with $Gender = `female'$ fall in the leaf labeled *Leaf1*, regardless of their value of *Age*. Units with $Gender = `male'$ and $Age \leq 35$ fall in the leaf labeled *Leaf2*, and units with $Gender = `male'$ and $Age > 35$ fall in the leaf labeled *Leaf3*.

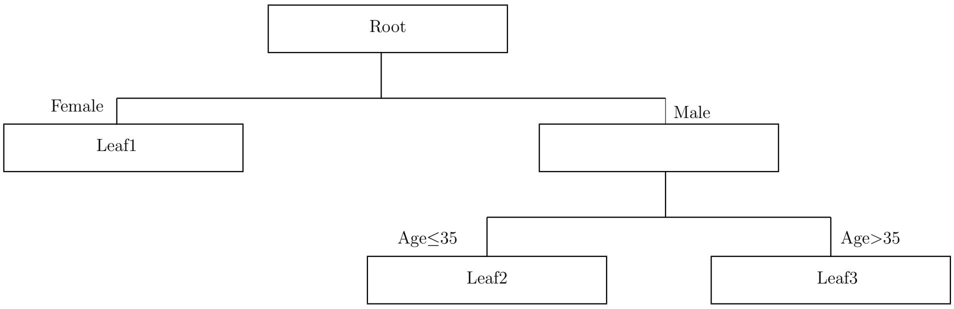


Figure 4.1: Example of a tree structure.

Within the *CART* method, an internal node is split by considering all allowable splits for all variables and the best split is that one with the most homogeneous daughter nodes. As thus, for the within-node impurity, two major criteria are required: A node splitting criterion to grow a large tree; and then a cost complexity criterion to prune a large tree. For instance, if the tree in Figure 4.1 is regarded too large or too complex, the branch to the leaves *Leaf2* and *Leaf3* can be cut, so that the resulting tree has only two leaves, *Leaf1* and what was formerly the root of *Leaf2* and *Leaf3*. Pruned trees typically do not predict the values in the observed data as well as larger ones, but they may be more robust to over-fitting than larger ones. This is the case because, pruning the tree too much may result in non-homogeneous imputation donors, so that the imputations are not drawn from plausible conditional distributions; essentially, the imputation classes are too broad. On the other hand, insufficiently pruning the tree may lead to over-fitting the observed data, resulting possibly in inferences with larger variances.

For a continuous response, as is the case in this application, a common choice of node impurity for node τ is the within-node variance of the response:

$$i(\tau) = \sum_{subject_i \epsilon \tau} (Y_i - \bar{Y}(\tau)) \tag{4.1}$$

where $\bar{Y}$ is the average of $Y_i^\prime s$ within node τ. To split a node τ into its two daughter nodes (the left and right) τ_L and τ_R the split function below is maximized

$$\phi(s,t) = i(\tau) - i(\tau_L) - i(\tau_R) \tag{4.2}$$

where s is an allowable split. The cost-complexity measure $R_\alpha(T)$ is defined as

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{4.3}$$

where $\alpha(\geq 0)$ is the complexity parameter, $|\tilde{T}|$the number of terminal nodes in $T$and the tree cost

$$R(T) = \sum_{\tau \in \tilde{T}} (i(\tau)) \tag{4.4}$$

Therefore, to select a right sized tree, minimal cost complexity pruning then involves identifying the unique smallest minimizing subtree $T(\alpha)$ for complexity parameter α. Breiman et al. (1984) suggests that an optimal sized tree may be considered as the smallest tree with a cross-validated misclassification error rate within one standard error of the tree with the minimum error rate. For a more detailed description of the method, readers are referred to (Breiman et al., 1984).

In what follows in this research, use of a *CART* refers to use of Regression trees since the response variables in focus are continuous.

Travel attributes

Travel-demand models assume that there are relationships between socio-demographic characteristics and travel characteristics, by using variables describing households/individuals, particularly in trip generation and mode-choice modeling. It is well founded that total trip generation is associated with the demographic and socio-economic attributes of the traveler (Ortzar and Willumsen, 2001). In the current study, we mainly model two variables. The first dependent variable used is the total number of trips made per day by individual respondents, which is conditioned on the travel purpose, the second is trip rates conditional on travel mode and lastly, travel duration for work trips.

Activity groups tend to be grouped differently by different researchers. In the past, activities were frequently divided into two types: work and leisure. The two-way classification has been used in several studies including activity-based trip generation modeling (Supernak et al., 1983; Munshi, 1993). However, consumer theory (Reichman, 1977; Lane and Lindquist, 1988) typically uses a three-way classification of activities into: (1) subsistence (income-producing or paid time, such as work); (2) nondiscretionary (obligated, maintenance or compulsory activities, such as eating meals, certain shopping, and child care); and (3) discretionary or leisure activities. Other researchers (Golob and McNally, 1997; Golob, 1998) use this classification in modeling relationships between activity and travel time. In this current study, a classification of the travel purpose attribute closely related to the latter 3-group classification is employed for the trip rates conditional on travel purpose goal. The independent variables used are the socio-demographic variables initially used as weighting variables in the data combination procedure: gender, age group, marital status and education level. Conditional on the attribute of interest, the homogeneous categories of the trip rates of individuals are developed using the regression tree methodology. Optimal sized trees are selected using the cross validation method.

4.2.2 Developing distributions

The next step towards setting up the simulation is developing distributions from which sampling can be conducted for the travel characteristic of interest using a Monte Carlo approach. An assessment was made on the shape and type of the distribution so as to select an appropriate distribution to simulate response values. Different distributions may be assumed for different attributes. In our case, for example, the Poisson distribution is assumed here in simulation of trip rates. This distribution was judged the most suitable due to the count nature of the response variables. This then provides the basis for the random sampling process used in the data simulation procedure.

The Poisson distribution is a discrete distribution. By definition, Y follows a Poisson distribution with parameter μ if and only if

$$P(Y = k) = \frac{\exp^{-\mu} \mu^k}{k!} \tag{4.5}$$

for $k = 0, 1, 2, \ldots$ with $E[Y] = \mu$ and $var[Y] = \mu$.

Following the developed distributions, the subsequent task was to sample from these distributions in order to simulate response values for the target sample. The data available here are randomly split by homogenous group into two sets: The training set, which is 75% of these data and the testing set being the remaining 25% of the data. While in principle, completely arbitrary proportions can be used, the percentages of 25%-75% are common practice in validation studies (see *e.g.* Wets et al., 2000). Following the described methodology, the different steps are applied on the training set in order to implement the simulation procedure. The testing set is later used for validation purposes.

To compare different data sets, the Poisson regression model, which is also sometimes called a log-linear model, was used. A Poisson regression model with $p - 1$ predictor variables assumes the form:

$$log(E[Y_i]) = \beta_0 + \sum_{j=1}^{p-1} \beta_i X_{i,j} \tag{4.6}$$

or equivalently:

$$E[Y_i] = exp(\beta_0 + \sum_{j=1}^{p-1} \beta_i X_{i,j}) \tag{4.7}$$

where $E[Y_i]$ is the expected value of the i^{th} observation of the dependent variable, $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ and $X_{i,1}, X_{i,2}, \ldots, X_{i,p-1}$ the corresponding observations of the explanatory variables (McCullagh and Nelder, 1989; Agresti, 2002). A Poisson regression model assumes that the mean and variance of the presumed Poisson distributed variable are equal. When the variance is significantly higher than the mean, the problem of over-dispersion arises and in the reverse case; under-dispersion is encountered. Potential over-dispersion is accounted for by using the deviance as the dispersion parameter. In practice, if the estimate of dispersion after fitting, as measured by the deviance or Pearsons chi-square, divided by the degrees of freedom, is not near 1, then the data may be overdispersed; if the dispersion estimate is greater than 1 or underdispersed; if the dispersion estimate is less than 1. A simple way to model this situation is to allow the variance functions of these distributions to have a multiplicative overdispersion factor ϕ. The variance function for the Poisson regression is then

$$V(\mu) = \phi(\mu) \tag{4.8}$$

An alternative method to allow for overdispersion in the Poisson distribution is to fit a negative binomial distribution in which case,

$$V(\mu) = \mu + k\mu^2 \tag{4.9}$$

instead of the Poisson. The parameter k can be estimated by maximum likelihood, thus allowing for overdispersion of a specific form. This is different from the multiplicative overdispersion factor ϕ, which can accommodate many forms of overdispersion (SAS Institute Inc., 2008).

4.2.3 An alternative simulation procedure

The approach proposed by Stopher et al. (2003) utilized *CART* to develop grouping used as the basis of the simulation. It has been recognized that *CART* are unstable; a small perturbation in the input variables or a fresh sample can lead to a very different classification tree. Some approaches exist that try to correct this instability. However, their benefits can, at present, be appreciated only qualitatively. The issue of instability of *CART* of course poses a difficulty. Replication of a given set of grouping produced by *CART* would be presumably difficult, even on the very same data set. To handle this limitation, we propose an alternative simulation method in this study. In the first step of categorizing individuals, the new method involves specifying a priori, a fixed set of socio-demographic groups based on all the variables of interest in the data. This is as opposed to Stopher et al. (2003) approach where the groups are determined by *CART*. It is proposed that the a priori specification of the groups be guided by the joint distribution of the variables of interest. In the next step of the new method, data are then simulated within each of these groups with respect to the response of interest using a Monte Carlo method. Since a separate survey data set was not used for validation, the available data were split into training and testing data sets. To maintain representative balanced samples within both the training and testing sets, splits were randomly made within groups. The data which are simulated by group were then merged to form a complete simulated data set. To investigate the robustness of the procedure to outliers, the training-testing splits were performed multiple times. Consequently, the simulations were also conducted multiple times and comparison between data sets is conducted. It is easy and possible to compare groups for different training-testing sets since these groups are similar each time. Moreover, at each training-testing split, it is also evaluated if results vary from simulation to simulation. The results obtained with the new method are compared with those from the above procedure. The shortcoming of this proposed method is the possibility of having few or no observations within a given specified group. The solution suggested here is to combine some groups so that sufficient observations exist within each socio-demographic group.

4.2.4 Comparison of results

The Poisson regression model is used to compare different sets of data. The Poisson regression together with other similar models such as the Negative Binomial and the Zero Inflated Poisson regression models are parametric methods that have also been commonly used in literature for the analysis of travel behavior. Amongst other applications, they have been used in comparison of pedestrian trip generation models (Kim and Susilo, 2008). Moreover, Targa and Clifton (2005) applied Poisson regression models to analyze built environment and nonmotorized travel.

Following the simulations, some basic descriptive statistics are initially provided to offer a comparison of the results from the simulated data with those from the actual survey data. In a later step, the Poisson regression model is fit on the data for further comparison. The Poisson regression model is also utilized to check for significant differences between distributions with the Wald test as the underlying test. For this test, within the generalized linear model, the Poisson distribution is specified and tests are conducted between the different groups of data with respect to the response. The null hypothesis is that the distributions of the testing data set and the simulated data set differ by a location shift of zero (hence no difference) and the alternative is that they differ by some other location shift.

Through the assessment of significant differences between the simulated data and the real travel survey data it is determined whether acceptable synthetic data, which would be useful for travel demand modeling can be generated through the simulation process.

The categories of the socio-demographic variables considered include: gender (male (1), female (2)), age (16-34 (1), 35-54 (2) and 55-75 (3) years of age), marital status (married (1), divorced (2), widowed (3) and unmarried (4)) and education level (primary school (1), junior high school (2), high school (3) and college or university (4)). The results obtained from application of the simulation procedure are discussed below and they follow three steps; categorizing individuals, data simulation and analyses to compare the simulated and actual survey data. The attributes of interest include trip rates and travel duration.

4.3 Trip rates by travel purpose

Conditional on the purpose for which individuals travel, interest is focused on simulation of trip rates in this Section. In this case, the 'purpose of travel' attribute is re-classified into three groups:

- the Subsistence group comprising of work-related trips, business visits and following education;
- the Maintenance group comprising of shopping trips, picking/dropping someone and trips for conducting services such as visiting the doctor, bank, among others;
- and lastly, Out-of-home leisure which involves visiting someone, trips due to sports/cultural/relaxation activities and also walking around.

4.3.1 Categorizing individuals

Under here, the main goal is to develop categorization schemes Regression trees, to predict the trip rates, whereby, different categorization schemes are developed for trip rates conditional on individual purpose of travel.

Overall, the results obtained show that the majority of the trips are performed due to maintenance (40%), followed by an approximately equal share of out-of-home leisure (30%) and subsistence (30%) trips. Regression tree-runs are initially completed for the three travel purposes.

Looking at the Subsistence group, it is observed that 6 homogeneous categories are obtained from the Regression tree procedure in the analysis of trip rates per person per day. Figure 4.2 shows the output of a regression tree-run obtained for Subsistence trips using the training data set. The cell means, the standard deviations together with the cell sizes are displayed within each terminal node. An optimal sized tree is obtained at the end of the Regression tree procedure, producing a final regression tree containing 6 terminal nodes. In Figure 4.2, it can be observed that all the four classification variables are used in the actual construction of the tree. The gender and education level of individuals are the top two variables to be split upon. This is a clear indication that these variables are of high importance for subsistence trips. The mean trip rates for the final categories range from 1.08 for male participants within the age group of *35-54* but with not more than high school level of education who are divorced or un-married to 1.81 for male participants with the highest level of education within the age group of *35-54* years (see also Table 4.1).

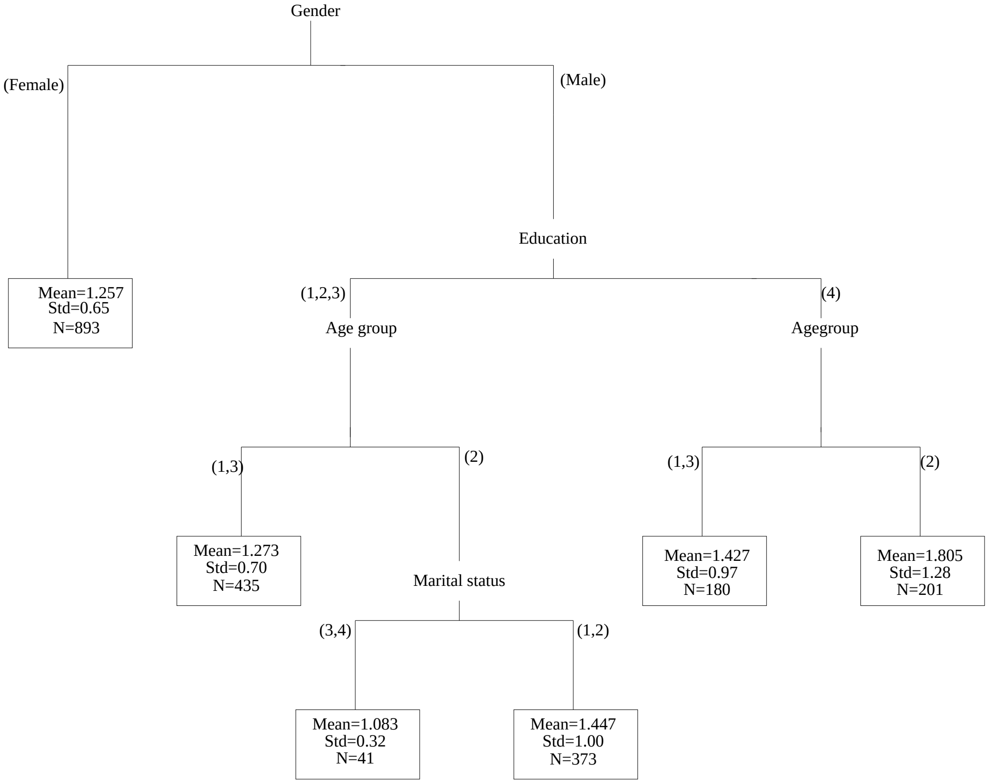


Figure 4.2: Regression tree segmentation results for the subsistence travel purpose trips.

Regarding the maintenance trips and out-of-home leisure trips, the regression tree segmentation results also into 6 and 5 homogeneous groups respectively. The final results of the categorizations scheme are shown in the Regression tree diagrams in Figure 4.3 and Figure 4.4. The results are also summarized in column 2 of Table 4.3 and 4.4 respectively, for the two trip purposes. As indicated in the displayed results, the classification of maintenance trips is highly dependent on the education level of individuals whereas the most important factor for out-of-home leisure trips is age group followed by education level.

At this point in the simulation procedure, the categorization of individuals has been accomplished. Based on the established categories, the next phase of the simulation procedure is to develop frequency distributions from which to sample, for the trip rates. Establishing a suitable distribution to be assumed is key here. We explored the plots of the probability distributions of the different groups with the Poisson distribution plots from corresponding means superimposed. The investigation showed that a Poisson distribution is a reasonable assumption in the next step of actual data simulation. The Wald test statistic was also computed to test for differences between distributions to confirm the visual impressions. For example for group 1 of the subsistence purpose and the corresponding Poisson distribution, the test result (p-value=0.8906) suggests no significant statistical differences. This was examined for several of the different homogeneous groups within the three travel purposes, and every time, using a Poisson distribution seemed a valid choice for simulation. Thus the distributions can be assumed to follow a Poisson distribution.

4.3.2 The simulation results

The next step in the simulation procedure is to perform the actual data simulation. To implement this, random samples are taken from the corresponding Poisson distributions for each created homogeneous category. For each distribution, a sample proportional to the size of the category in the testing data set is initially drawn and is compared to the testing data set. This draw is then conducted multiple times, in this case 5 times and the simulation results are combined for further comparison. The purpose of the multiple draws is to check for possible differences that may arise between simulations. Consequently, a travel survey data set for each individual can be synthetically generated with regards to the purpose of travel trip rates. Table 4.1 gives a comparison of the subsistence trip rates for the testing data set and the simulated data following the earlier generated homogeneous categories. For completeness, the results corresponding to the training data set are also shown. The second column results, detail the categorization scheme for each of the six groups. Results are also shown for the data that are simulated only once in the second last column and for the data that were simulated multiple times in the last column.

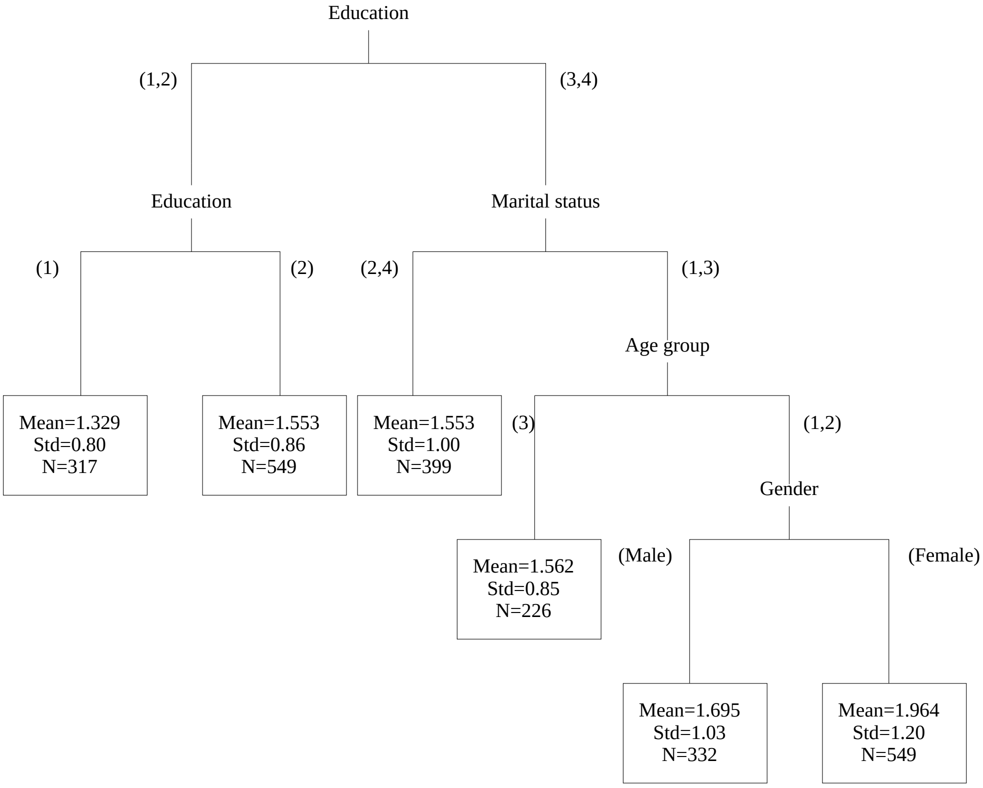


Figure 4.3: *CART* segmentation results for the maintenance travel purpose trips.

Using the Wald test, for the comparison between the data that is simulated (once as well as 5 times) and the testing data set, no statistical significant differences exist between the data sets for all groups (for example, for group 1: z-value=-0.187, p-value=0.8519, group 4: z-value=1.487, p-value=0.1370 and group 6: z-value=0.707, p-value=0.4796). No overall statistical differences were also found between these data (z-value=0.545, p-value=0.5856). However if another test, the Wilcoxon rank sum test is applied, no statistical differences exist for groups 1, 2, 3 and 5 but an overall significant difference ($W = 227996$, p-value $= 0.00015$) is observed between the simulated and the travel survey data. Significant differences were also found between the data sets (testing and simulated data) for group 4 ($W = 5362$, p-value $= 0.00047$) and group 6 ($W = 2300$, p-value $= 0.00196$) for this test. More so, simulating these data multiple times may not necessarily provide results that are closer to

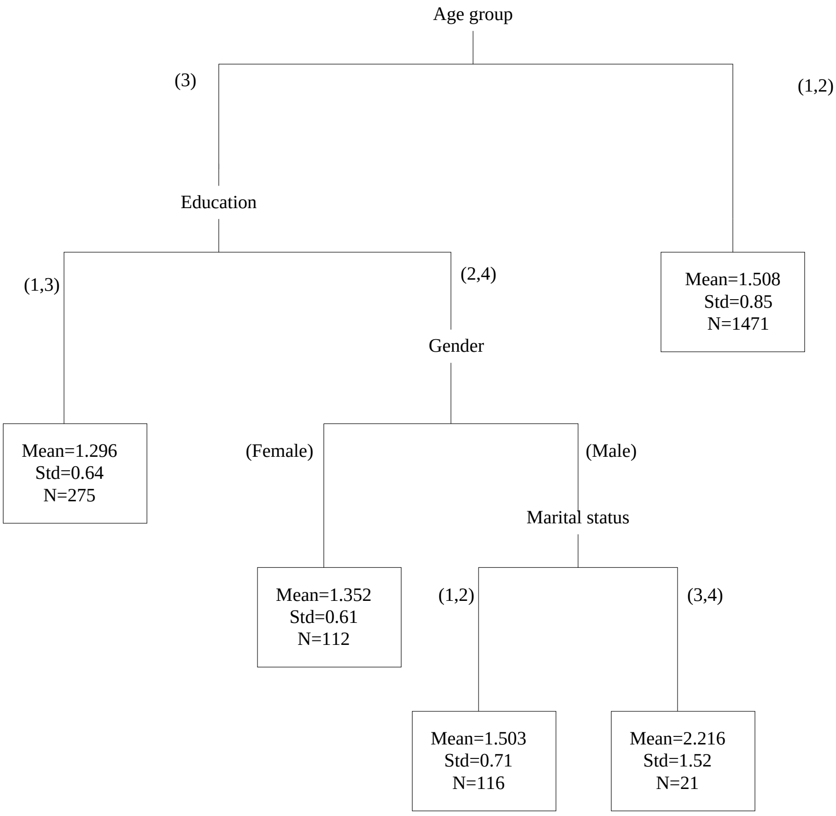


Figure 4.4: *CART* segmentation results for the out-of-home leisure travel purpose trips.

the testing set as evidenced by the results for the data simulated 5 times. For groups where significant differences are found, what would be of further interest is to check for the possible initial differences between the training and testing data sets. Indeed, significant differences do exist between these sets of data for groups 4 and 6 using the Wilcoxon test. However, interestingly, as noted before, all these reported differences are not significant when the Wald test is used. Use of the Wald test is beneficial due to the fact that the known underlying Poisson distribution is specified and not ignored. Furthermore, the test is suitable since the sample sizes are relatively large in the available data. The Wilcoxon test tends to find significant differences when sample sizes are large, as is the case here, and thus a conclusion based on this test would be considered to be the worst case scenario in this respect. From now on, only results of the Wald test are reported.

The simulated trip rates per person per day range from 1 to 5 trips, which is a narrower margin in comparison with both the training and the testing data sets that range from 1 to 9 trips. However, this is not of much concern as only a few outlying individuals with trips rates greater than 5 in the actual survey data were identified. The mode, which is noted to be a more robust central measure as compared to the mean with regards to trip rates, is also obtained. It is revealed that the mode is equal to 1 for the simulated data as well as the testing data set (and training data set) with respect to each homogeneous group. From all the described findings, it can be said that, the subsistence trip rates of the actual travel survey data seem to be generally replicated relatively well in the simulated data across the groups.

Table 4.3 shows a comparison of the simulated maintenance trip rates with those from the actual survey data (the testing and the training data set), following the corresponding developed categorization scheme for the maintenance group. The simulated and the testing data appear to compare favorably across the individual groups and overall since no statistical significant differences are found using the Wald test statistics at the 5% level of significance. The simulated maintenance trip rates range from 1 to 6 whereas those from the testing data set are from 1 to 7. As in the Subsistence trips case, the mode is maintained at 1 for both the simulated and testing data sets for each of the groups.

Focusing on the out-of-home leisure trip rates, the corresponding results as shown in Table 4.4, highlight a comparison of the simulated trip rates with those from the actual survey data (the testing and the training data set), following the respective developed categorization scheme for the out-of-home leisure group. For the data that are simulated once, again no overall significant differences are observed between these data and the testing set (z-value=0.307, p-value=0.759) and no significant differences exist between groups.

Table 4.1: Comparison of the subsistence (travel purpose) trip rates between the survey data and the simulated data

		Mean (Standard deviation)			
Group	Categorization Scheme[a]	Training Set	Testing Set	Simulated	Simulated
		(n=2123)	(n=708)	Data (once)	Data (5 times)
1	Gender=female	1.26(0.65)	1.25(0.64)	1.25(0.49)	1.27(0.52)
2	Gender=male, Education =1,2,3, Age group =1,3	1.27(0.70)	1.28(0.71)	1.29(0.56)	1.29(0.54)
3	Gender=male, Education =1,2,3, Age group =2, Marital status=3,4	1.08(0.32)	1.15(0.40)	1.05(0.22)	1.07(0.26)
4	Gender=male, Education =1,2,3, Age group =2, Marital status=1,2	1.45(1.00)	1.29(0.62)	1.52(0.66)*	1.46(0.72)*
5	Gender=male, Education =4, Age group =1,3	1.43(0.97)	1.61(1.54)	1.46(0.66)	1.46(0.71)
6	Gender=male, Education =4, Age group =2	1.81(1.28)	1.72(1.30)	1.90(0.76)*	1.85(0.85)*
	Overall	1.35(0.85)	1.33(0.86)	1.39(0.62)*	1.38(0.64)*
Trip-range		[1,9]	[1,9]	[1,4]	[1,5]

[a] Category codes are as defined in Section 4.2.4, *Significantly different from the corresponding testing data set result using the Wilcoxon rank sum test

Table 4.2: Comparison of the subsistence (travel purpose) trip rates between the survey data and the *FIX* simulation procedure data

Group	Categorization Scheme[a]	Mean (Standard deviation)		
		Training Set	Testing Set	Simulated Data
		(n=2118)	(n=712)	(n=712)
1	Gender=female	1.24(0.66)	1.29(0.54)	1.31(0.71)
2	Gender=male, Education =1,2,3, Age group =1,3	1.27(0.64)	1.27(0.54)	1.26(0.76)
3	Gender=male, Education =1,2,3, Age group =2, Marital status=3,4	1.12(0.33)	1.20(0.42)	1.10(0.32)
4	Gender=male, Education =1,2,3, Age group =2, Marital status=1,2	1.44(1.03)	1.60(0.85)	1.36(0.81)
5	Gender=male, Education =4, Age group =1,3	1.48(1.15)	1.32(0.50)	1.41(1.17)
6	Gender=male, Education =4, Age group =2	1.75(1.36)	1.52(0.73)	1.90(1.59)
	Overall	1.35(0.87)	1.38(0.93)	1.52(0.73)
Trip-range		[1,9]	[1,9]	[1,7]

[a] Category codes are as defined in Section 4.2.4

Table 4.3: Comparison of the maintenance (travel purpose) trip rates between the survey data and the simulated data

Group	Categorization Scheme[a]	Mean (Standard deviation)			
		Training Set	Testing Set	Simulated	Simulated
		(n=2123)	(n=708)	Data (once)	Data (5 times)
Group	Categorization Scheme[a]	Training Set	Testing Set	Simulated	Simulated
		(n=2372)	(n=720)	Data (once)	Data (5 times)
1	Education =1	1.33(0.80)	1.31(0.92)	1.32(0.51)	1.35(0.56)
2	Education =2	1.55(0.86)	1.52(0.83)	1.54(0.78)	1.61(0.78)
3	Education =3,4, Marital status=2,4	1.56(1.00)	1.59(1.06)	1.61(0.77)	1.55(0.73)
4	Education =3,4, Marital status=1,3, Age group =3	1.56(0.85)	1.59(0.91)	1.51(0.69)	1.54(0.72)
5	Education =3,4, Marital status=1,3, Age group =1,2, Gender=male	1.69(1.03)	1.75(1.09)	1.66(0.87)	1.64(0.80)
6	Education =3,4, Marital status=1,3, Age group =1,2, Gender=female	1.96(1.20)	1.93(1.20)	2.17(1.03)	2.02(1.02)
	Overall	1.62(1.01)	1.62(1.03)	1.64(0.87)	1.64(0.82)
Trip-range		[1,9]	[1,7]	[1,6]	[1,6]

[a] Category codes are as defined in Section 4.2.4

Table 4.4: Comparison of the out-of home leisure (travel purpose) trip rates between the survey data and the simulated data

Group	Categorization Scheme[a]	Mean (Standard deviation)			
		Training Set (n=1995)	Testing Set (n=665)	Simulated Data (once)	Simulated Data (5 times)
1	Age group =3, Education =1,3	1.30(0.64)	1.36(0.68)	1.30(0.60)	1.31(0.58)
2	Age group =3, Education =2,4, Gender=female	1.35(0.61)	1.37(0.52)	1.29(0.46)	1.36(0.54)
3	Age group =3, Education =2,4, Gender=male, Marital status=1,2	1.50(0.71)	1.20(0.45)	1.60(0.76)	1.56(0.80)
4	Age group =3, Education =2,4, Gender=male, Marital status=3,4	2.22(1.52)	1.25(0.44)	1.71(0.76)	2.37(1.09)
5	Age group =1,2	1.51(0.85)	1.52(0.78)	1.55(0.72)	1.52(0.71)
	Overall	1.47(0.82)	1.48(0.74)	1.50(0.70)	1.49 (0.71)
Trip-range		[1,7]	[1,5]	[1,4]	[1,5]

[a] Category codes are as defined in Section 4.2.4

Taking an overall view on the results, on average, concerning both the simulated and actual travel survey data, maintenance trip rates per person per day are observed to be the highest followed by out-of-home leisure trip rates with the subsistence trip rates being the least. Some significant differences between groups have been indicated in the results when comparing the simulated with the actual data using the Wilcoxon test. This finding is not so surprising in such an experiment with varying group sizes. Such a difference may be attributable to small sample sizes of groups and the high variability that may exist within some groups. Much as training here is based on a larger sample of the group, testing the data on some minority groups tends to be sensitive to outliers. However, since no statistical differences were observed in this study for all groups using the Wald test, there is no concern and it can be argued that there is evidence to believe that this method works well and may be prospective for future applications.

A question that may be interesting to discuss here with regard to this experiment would be: How small should the sample size of the original survey be to permit acceptable results from this simulation method? This, of course, is not a question that can have one precise answer. The minimum estimation sample size may vary from study to study as it depends on different factors. These factors include the number of socio-demographic and economic variables considered in a study and the resultant number of Regression tree segmentation groups. In our opinion, if one is able to achieve a 'reasonable' number of units within each homogeneous group generated by the Regression tree, then the simulation approach should still yield acceptable results. Again here, the 'reasonable' number of units within each homogeneous group may be defined differently by different researchers. Some authors have proposed a minimum number of 5 within each node (Breiman et al., 1984). All being said, it is perhaps most reasonable to argue that, research based on larger sample sizes that are representative of the target population are more preferable. This facilitates improvement of the robustness of the method, since the method involves sub-dividing the data into several groups. It should be interesting however, to investigate the influence of varying sample sizes in application of this simulation method. This could be achieved by setting several experiments incorporating different survey sample sizes. We do not dwell on this any further in this thesis, but it may be a possible area for future exploration.

To examine robustness and stability of the applied methodology, additional multiple experiments were conducted here. The aspect of robustness and stability is often not researched in travel simulation studies. It was of interest here, to check for differences between different splits of the training and testing sets. This is important since one could argue that the split performed here arose by chance and hence the resultant procedure conducted on the training set to simulate data may also have yielded the observed results by chance. It was of interest to check for differences between different testing sets versus the different simulated data so as to give an indication of the variability between simulations. Using Stopher et al. (2003), it would not be meaningful to conduct the entire experiment multiple times for the purpose of this comparison. This is due to the instability of *CART* as discussed earlier on. To conduct several comparisons while accommodating this limitation, we assume the segmentation groups generated in one experiment to be fixed. The main data set is randomly split (within each group) 1000 times into different training and testing sets. Thus, within each training-testing split, the fixed groups are constructed each time. Several simulations are consequently made within each group based on the training sets. At each split, tests for differences between the training and testing set are conducted with further tests to check for differences between the testing and the simulated data. Overall, in 1000 splits, for the subsistence travel purpose, only about 0.3% significant differences were found between the training and testing sets for the overall mean trip rates. In testing for differences between the simulated and testing data, 0.6% differences are found. Similar results were obtained with respect to other travel purposes. The results demonstrate the robustness of the method to outliers and evidence against chance findings. Furthermore, support is provided for the stability of results from different simulation runs. This is an interesting finding that as not been investigated before to the best of our knowledge.

4.3.3 Alternative simulation procedure

Let us now consider the results of the alternative simulation procedure that we propose in this study. To define a fixed set of demographic groups a priori, for this alternative simulation procedure(here after referred to as *FIX*), the joint distribution of the available socio-demographic variables was exploited.

The variables were re-categorized to ensure sufficient units in each of the subsequent fixed groups. These variables; gender (male or female), age (16-34 or 35-75 years of age), marital status (married or divorced/widowed/unmarried) and education level (primary school/junior high school or high school/college/university) thus resulted into 16 predetermined groupings coming from the multi-way cross classification. Table 4.2 shows the results of the *FIX* procedure for the subsistence travel purpose. For comparison purposes, the resultant simulated data are grouped in the same way as before[see Table 4.1], and the results are provided by group. Despite the fact that these were not the original groups on which the simulations were developed, it can be observed that the results of this alternative simulation procedure were quite similar to those obtained above. Of course, there is the possibility that one of the advantages working in favor of our method could be the fact that more groups were used than in the simulation following Stopher et al. (2003) approach. Nevertheless, it could be concluded that the simulated data replicate the actual data reasonably well overall and with respect to each fixed group since no significant differences were observed. An added value of our method is that it is easier to implement. Moreover, the simplicity of the approach makes it also easier to conduct multiple simulations as the groups are similar by each time. If simulations were required for several demographic regions, the method would be easier to implement with minimal human intervention to determine the necessary number of groups.

As far as stability of results is concerned, in 1000 splits, about 0.2% significant differences found overall between the training and the testing data for the subsistence purpose. There were about 0.4% statistically significant differences between the simulated and the testing data in 1000 splits. This result is better than what was observed under Stopher et al. (2003), which further underscores our method as a favorable approach. However, it would still be interesting to apply this method in other settings to confirm this result and to investigate if the result can be generalized. Stopher et al. (2003, 2005); Mohammadian et al. (2010) argue that a classification scheme is critical in simulating household travel survey data, and that the more homogeneous groups the population and the dependent variables are subdivided into, the better the chance will ultimately be for the simulation to reflect these differences. For classification, Stopher et al. (2003, 2005) rely on *CART*.

Mohammadian and Zhang (2007) also uses a clustering approach in an investigation on transferability of national household travel survey data. This study has established with respect to this application, that a data simulation procedure that is based on an a priori classification scheme can also perform favorably well.

4.3.4 Results of the Poisson Regression Model

To investigate the level of precision that can be achieved with the simulation model utilized, the Poisson regression model was separately fit to both the testing and simulated sets of data with respect to the different travel purposes. The data that are a result of a single simulation are utilized in these analyses. The model is also fit on the training data set. However, since the training data set was used as the basis of the simulation, comparison with this set is of less interest. At the start of the model building process, the full the model is initially fit based on all four socio-demographic variables (gender, education level, marital status and age group) together with all possible two-way interactions between the variables. This is done With respect to each data set. Overall, in the model building process, some problems of possible underdispersion were observed when the Poisson regression model was fit on the data. Corrective measures taken involved using the deviance as an estimate of the dispersion parameter instead of setting it to 1.

For the subsistence purpose, in the model building process all interactions together with the main effects of marital status and age group are found to be statistically insignificant and therefore dropped from the model. The final model, which is also the same model obtained from all three sets of data, comprises of only gender and education level. Linking back to the simulation procedure, these are also the variables initially found to be of high importance in the regression tree results. In this way, the Poisson regression model validates the Regression tree groupings, but it also adds an extra dimension here in supporting/validating the underlying procedure that was used to generate the analyzed simulated data. Table 4.5 shows the results of the final Poisson regression model in modeling the subsistence trip-rates for each of the data sets. Again, for completeness, the results corresponding to the training data set are shown as well. The results for the testing and the simulated data are quite close. Impressively, the parameter estimates are generally consistent in direction and magnitude, with similar estimates of standard errors. More to this, the ranges of the confidence intervals for the respective parameters are relatively similar, with a significant overlap.

Table 4.5: Modeling the subsistence trip rates using the Poisson regression model with respect to the training, testing and simulated data sets

Data set	Effect	Estimate(s.e)	95% CI	p-value
Training	Intercept	0.31(0.024)	[0.26, 0.35]	<.0001
	Male	0.14(0.023)	[0.09, 0.18]	<.0001
	Primary school	-0.22(0.052)	[-0.33, -0.12]	<.0001
	Junior high school	-0.12(0.031)	[-0.18, -0.06]	<.0001
	High school	-0.11(0.027)	[-0.16, -0.06]	<.0001
Testing	Intercept	0.38(0.039)	[0.30, 0.45]	<.0001
	Male	0.12(0.040)	[-0.05, 0.20]	0.0021
	Primary school	-0.33(0.095)	[-0.52, -0.15]	0.0005
	Junior high school	-0.29(0.054)	[-0.39, -0.18]	<.0001
	High school	-0.21(0.045)	[-0.29, -0.12]	<.0001
Simulated	Intercept	0.33(0.031)	[0.27, 0.39]	<.0001
	Male	0.18(0.031)	[0.12, 0.24]	<.0001
	Primary school	-0.28(0.073)	[-0.42, 0.13]	0.0001
	Junior high school	-0.19(0.041)	[-0.27, -0.11]	<.0001
	High school	-0.16(0.035)	[-0.23, -0.09]	<.0001

Table 4.6 shows the results corresponding to the maintenance travel purpose. The model building process yielded a main effects model including all variables for the training data set. However, for the testing and simulated data, the effects related to marital status were not significant. Nevertheless, we kept this variable in the models to maintain an overall picture. Differences in parameter estimates for the testing versus the simulated data set in terms of direction and magnitude are observed only in the *Widowed* effect. However, for the same effect, the corresponding confidence intervals for the parameter still cover quite a closely overlapping range. Moreover, it is to be noted that the two sets of confidence intervals cover zero. The confidence intervals for the rest of the parameter estimates are considerably similar.

Table 4.7 offers a display of the results corresponding to the out-of-home leisure travel purpose. For all data sets, the results show that all the variables are found to be statistically insignificant with the exception of *age*. The confidence intervals of age are acceptably overlapping. It is also observable that other parameter estimates with respect to the testing and simulated data are approximately consistent in direction and magnitude with the except for the *Primary school*, *Married* and *Widowed* effects. Without over-interpreting the results, it can be said that there is not so much more to learn from this model.

In the interest of getting a feel of what the data has to offer on an aggregated level, the Poisson regression was fitted to the complete synthetic data set without splitting on travel purpose. Table 4.8 shows the results of the final fitted model, after model building. The model contains only age and education as the important variables in explaining trip rates. This indicates that if we do not consider the purpose for which people travel, other factors remaining constant, age and education level are the important variables in explaining the variability in trips made by individuals. It is demonstrated that consistence of parameter estimates in magnitude and direction is maintained as observed in the earlier models.

Most of the studies on simulation techniques have concentrated on descriptive methods in evaluating the quality of synthetic data. For example, in Stopher et al. (2003) to determine acceptable synthetic data, the investigation aimed at producing numbers that are reasonably close to those that would have been obtained from an actual survey. In their earlier research (Greaves and Stopher, 2000), the primary measure of acceptability used was statistical significance of differences between the simulated data and a real household travel survey. However, it is notable that statistically significant differences can be found when numerical differences are very small. Moreover, statistical significance of differences may be regarded by some researchers as not being a necessary nor a sufficient measure of acceptability. The debate on what should be preserved through simulation and how the quality of simulated data should be evaluated, has been active and is likely to remain open for the coming several of years as the field grows. These debates have been hosted mainly during different international conferences that have had sessions on simulated or synthetic data. Whereas some schools of thought advocate for preservation marginal distributions others argue that preservation of statistics such as the mean, mode, is the key issue.

Our investigation has gone further to employ a standard statistical model to evaluate the quality of synthetic data generated. We have also investigated the robustness and stability of the simulation methodology by setting up multiple experiments. This is a new finding that adds great value to the developing field of travel data simulation. From the fitted models, it seems that the parameters are estimated fairly comparatively across the data sets; simulated and 'true' data. Overall, although we do not claim that simulated data are the perfect basis for analysis of travel behavior, they do provide results that are similar to those produced from actual travel survey data. The findings by (Stopher et al., 2003, 2005, 2007) also provide strong support for data simulation. They establish that simulating household travel characteristics using the approaches that they proposed, produces reasonable approximations to actual travel characteristics obtained from *HTSs.* Furthermore, very good fit has been established between the means of the simulated and the validation data, both for trips per person and trip distance per person (Zhang and Mohammadian, 2008b). Based on the findings, it seems promising to use the travel data simulation approaches. However, when employing travel data simulation approaches in settings where this is required, we recommend use of large, quality samples as the basis for the simulations as this is important in determining the quality of the resultant synthetic data.

Table 4.6: Modeling the maintenance trip rates using the Poisson regression model with respect to the training, testing and simulated data sets

Data set	Effect	Estimate(s.e)	95% CI	*p*-value
Training	Intercept	0.41(0.048)	[0.32, 0.50]	<.0001
	Male	-0.07(0.024)	[-0.12, -0.03]	0.0018
	Primary school	-0.27(0.042)	[-0.35, -0.19]	<.0001
	Junior high school	-0.15(0.032)	[-0.21, -0.08]	<.0001
	High school	-0.10(0.028)	[-0.16, -0.05]	0.0003
	16-34 years	0.14(0.038)	[0.06, 0.21]	0.0003
	35-54 years	0.10(0.031)	[0.04, 0.16]	0.0009
	Married	0.17(0.034)	[0.10, 0.24]	<.0001
	Divorced	0.04(0.066)	[-0.09, 0.17]	0.5727
	Widowed	0.13(0.074)	[-0.01, 0.28]	0.0732
Testing	Intercept	0.40(0.080)	[0.25, 0.56]	<.0001
	Male	-0.06(0.041)	[-0.14, 0.02]	0.1383
	Primary school	-0.18(0.073)	[-0.32, -0.03]	0.0157
	Junior high school	-0.14(0.057)	[-0.26, 0.03]	0.0116
	High school	-0.05(0.048)	[-0.15, 0.04]	0.2741
	16-34 years	0.13(0.063)	[0.002, 0.25]	0.0463
	35-54 years	0.14(0.054)	[0.03, 0.24]	0.0112
	Married	0.13(0.059)	[-0.00, 0.24]	0.0352
	Divorced	0.01(0.117)	[-0.22, 0.24]	0.9103
	Widowed	-0.10(0.140)	[-0.37, 0.18]	0.4924
Simulated	Intercept	0.46(0.066)	[0.33, 0.59]	<.0001
	Male	-0.14(0.034)	[-0.21, -0.08]	<.0001
	Primary school	-0.29(0.062)	[-0.41, -0.17]	<.0001
	Junior high school	-0.16(0.047)	[-0.25, -0.07]	0.0006
	High school	-0.05(0.040)	[-0.13, 0.03]	0.2105
	16-34 years	0.13(0.052)	[-0.03, 0.23]	0.0139
	35-54 years	0.14(0.045)	[-0.05, 0.23]	0.0020
	Married	0.13(0.049)	[0.03, 0.22]	0.0105
	Divorced	0.09(0.094)	[-0.09, 0.27]	0.3374
	Widowed	0.10(0.108)	[-0.11, 0.32]	0.3339

Table 4.7: Modeling the Out-of-home Leisure trip rates using the Poisson regression model with respect to the training, testing and simulated data sets

Data set	Effect	Estimate(s.e)	95% CI	*p*-value
Training	Intercept	0.37(0.047)	[0.27, 0.46]	<.0001
	Male	0.02(0.023)	[-0.02, 0.07]	0.3577
	Primary school a	-0.07(0.043)	[-0.15, 0.02]	0.1200
	Junior high school	-0.05(0.031)	[-0.11, 0.01]	0.0992
	High school	-0.05(0.028)	[-0.10, 0.01]	0.0899
	16-34 years	0.09(0.038)	[-0.01, 0.17]	0.0197
	35-54 years	0.07(0.031)	[-0.01, 0.13]	0.0220
	Married	-0.02(0.033)	[-0.08, 0.05]	0.5556
	Divorced	-0.003(0.064)	[-0.12, 0.13]	0.9638
	Widowed	0.09(0.073)	[-0.05, 0.24]	0.2089
Testing	Intercept	0.26(0.073)	[0.12, 0.41]	0.0003
	Gender(Male)	0.03(0.036)	[-0.04, 0.10]	0.3949
	Primary school	-0.06(0.066)	[-0.19, 0.07]	0.3636
	Junior high school	-0.14(0.052)	[-0.24, -0.04]	0.0069
	High school	-0.06(0.043)	[-0.15, 0.02]	0.1451
	16-34 years	0.20(0.060)	[0.09, 0.32]	0.0007
	35-54 years	0.12(0.050)	[0.02, 0.22]	0.0154
	Married	0.08(0.050)	[-0.02, 0.17]	0.1307
	Divorced	-0.09(0.108)	[-0.30, 0.12]	0.4070
	Widowed	0.11(0.121)	[-0.12, 0.35]	0.3492
Simulated	Intercept	0.32(0.070)	[0.18, 0.46]	<.0001
	Gender(Male)	0.04(0.034)	[-0.03, 0.11]	0.2515
	Primary school	0.02(0.063)	[-0.11, 0.14]	0.8052
	Junior high school	-0.002(0.049)	[-0.10, 0.09]	0.9705
	High school	-0.02(0.042)	[-0.10, 0.07]	0.6936
	16-34 years	0.11(0.057)	[-0.01, 0.22]	0.0634
	35-54 years	0.13(0.047)	[0.04, 0.23]	0.0044
	Married	-0.03(0.048)	[-0.12, 0.07]	0.5577
	Divorced	-0.07(0.098)	[-0.26, 0.12]	0.4712
	Widowed	-0.05(0.118)	[-0.28, 0.18]	0.6604

Table 4.8: Modeling the All-purpose trip rates using the Poisson regression model with respect to the training, testing and simulated data sets

Data set	Effect	Estimate(s.e)	95% CI	*p*-value
Full data	Intercept	0.46(0.03)	[0.39, 0.52]	<.0001
	Primary school	-0.18(0.04)	[-0.26,-0.09]	<.0001
	Junior high school	-0.19(0.03)	[-0.25,-0.13]	<.0001
	High school	-0.10 (0.03)	[-0.16,-0.05]	0.0002
	16-34 years	0.02(0.03)	[-0.04, 0.09]	0.4770
	35-54 years	0.07(0.03)	[0.01, 0.13]	0.0326
Simulated	Intercept	0.44(0.03)	[0.38,0.49]	<.0001
	Primary school	-0.17(0.04)	[-0.25,-0.10]	<.0001
	Junior high school	-0.12(0.03)	[-0.17,-0.07]	<.0001
	High school	-0.07(0.02)	[-0.12,-0.03]	0.0022
	16-34 years	0.02(0.03)	[-0.04,0.08]	0.5303
	35-54 years	0.10(0.03)	[0.05,0.15]	0.0003

4.4 Trip rates by travel mode

When studying travel behavior, it is often interesting to understand peoples' preferential modes of travel as various modes of travel can be used to effect travel. In Flanders, car mode forms the highest share of trip rates of 64.5%, followed by use of slow mode, which accounts for about 23% of the total trips made. Slow mode is defined here to include travels on foot or by bicycle. The rest of the trip rates are conducted using other modes of travel such as public transport and motorbikes. Use of public transport for instance forms about 3.8% of the overall total trips and use of motorbikes forms about 1.8%. The 'other modes' of travel also incorporate use of other modes that were undefined by the users. As in Nakamya et al. (2007a), in this sub section, trip rates are simulated conditional on travel mode. The travel mode is grouped into three; car, slow mode and 'other mode' of travel. Regression tree runs are thus completed for the travelers or participants in travel by these modes of travel.

The Regression tree results obtained for car mode trips contain eleven terminal nodes thus forming eleven 'homogeneous' categories of the segmentation results. All the four classification variables used in the earlier analyses of travel purpose are again utilized in the construction of the tree. The education level and age group of individuals are the first two variables to be split upon, indicating their high importance for car mode trips. The mean trip rates for the final formulated categories in the terminal nodes range from 2.4 for persons in the 35-54 age group and with primary or junior high school education level to a mean trip rate close to 3.8, for persons in the age range of 16 to 54 years, who are either married or divorced and have a college/university degree. Basing on the established categories, in the next phase of the simulation procedure we develop frequency distributions from which samples are taken. As assumed earlier, the Poisson distribution is again the suitable choice for the distribution of trip rates. Moreover, exploration of the different frequency distributions superimposed onto a Poisson distribution supports that these trip counts can be assumed to follow a Poisson distribution. Noteworthy however is that, since the data set does not contain zero trips (only travelers are considered), a modified sampling approach is implemented such that only values other than zero are selected from the distribution. Random samples are taken from the corresponding distributions for each created 'homogeneous' category. Table 4.9 gives a comparison of the car mode trip rates for the combined survey data with the simulated data following the earlier generated categories. Thus, based on the categories created, summary statistics are also displayed using the *FHTS'00* data and are shown in the third column of Table 4.9. Similar conclusions are arrived at here as in the previous sub section. In this case, the car trip rates of the travel survey data seem to be replicated relatively well by the simulated data across the categories.

Considering the foot/bicycle trips, the Regression tree segmentation results into six homogeneous groups. These trips are highly dependent on the age group of individuals. The older persons (55-75 years) stand out with the minimum trip rates. If the individuals are between 16 and 54 years of age, the next important factor is their marital status. Table 4.10 shows a comparison of the bike/walk mode simulated trip rates with the trip rates from the survey data, following the corresponding developed categorization scheme. The foot/bicycle trip rates from the simulated data also seem to be relatively close to the results from the survey data (column 4 and 5), although slightly higher for most of the categories. Nevertheless, this could be by chance since some lower values are observed as well.

Considering the 'other' mode trips, five final categories are formed and the most important factor is found to be the individuals' marital status. In Table 4.11, the trip rates from the simulated data tend to display more variation from the survey data across the categories. This is could be highly attributable to the small cell sizes groups, on which the simulation is based and possibly also the high variances between trip rates.

Table 4.9: Comparison of the Car mode trip rates between the survey data and the simulated data

		Mean (Standard deviation)		
Group	Categorization Scheme[a]	*FHTS'00* data	Synthetic sample	Simulated Data
1	Education=1,2, Age group=3	2.43(1.22)	2.41(1.27)	2.61(1.36)
2	Education=1,2, Age group=1,2, Marital status=3,4	2.49(1.22)	2.53(1.32)	2.55(1.34)
3	Education=1,2, Age group=2, Marital status=1,2	2.92(1.61)	2.88(1.62)	3.08(1.62)
4	Education=1,2, Age group=1, Marital status=1,2	3.60(2.03)	3.58(1.99)	3.71(1.89)
5	Education=3, Age group=3	2.80(1.38)	2.75(1.37)	3.00(1.64)
6	Education=3, Age group=1,2, Marital status=2,4	3.06(1.95)	3.03(1.96)	3.22(1.59)
7	Education=3, Age group=1,2, Marital status=1,3, Gender=1	3.26(1.77)	3.24(1.70)	3.50(1.76)
8	Education=3, Age group=1,2, Marital status=1,3, Gender=2	3.60(1.94)	3.62(2.00)	3.58(1.86)
9	Education=4, Age group =3	3.40(1.75)	3.16(1.74)	3.02(1.73)
10	Education=4, Age group=1,2, Marital status=2,4	3.46(2.16)	3.38(2.03)	3.31(1.65)
11	Education=4, Age group=1,2, Marital status=1,3	3.85(1.99)	3.77(1.98)	3.79(1.84)
	Overall	3.17(1.82)	3.13(1.82)	3.27(1.72)

[a] Category codes are as defined in Section 4.2.4

Table 4.10: Comparison of the Bike/Walk mode trip rates between the survey data and the simulated data

		Mean(Standard deviation)		
Group	Categorization Scheme[a]	*FHTS'00*data	Synthetic sample	Simulated Data
1	Age group=3	2.22(1.13)	2.11(1.18)	2.33(1.29)
2	Age group=1,2, Marital status=1,3, Gender=1	2.17(1.16)	2.13(1.16)	2.35(1.30)
3	Age group=1,2, Marital status=1,3, Gender=2	2.35(1.40)	2.39(1.47)	2.67(1.44)
4	Age group=1,2, Marital status=2,4, Education =2,4	2.38(1.47)	2.38(1.45)	2.61(1.36)
5	Age group =1,2, Marital status=4, Education =1,3	2.55(1.54)	2.42(1.51)	2.78(1.57)
6	Age group=1,2, Marital status=2, Education =1,3	3.25(2.15)	3.22(1.93)	3.00(1.69)
	Overall	2.33(1.33)	2.28(1.35)	2.52(1.39)

[a] Category codes are as defined in Section 4.2.4

Table 4.11: Comparison of the ‘Other’ mode trip rates between the survey data and the simulated data

		Mean(Standard deviation)		
Group	Categorization Scheme[a]	*FHTS'00* data	Synthetic sample	Simulated Data
1	Marital status=1,3	1.90(0.94)	1.86(0.96)	2.15(1.19)
2	Marital status=2,4, Age group=2	1.98(1.07)	1.96(1.10)	2.16(1.13)
3	Marital status=2,4, Age group =1,3, Education =3,4	1.98(1.34)	2.02(1.31)	2.30(1.22)
4	Marital status=2,4, Age group =1,3, Education =1,2, Gender=1	2.07(1.22)	1.97(1.17)	2.55(1.36)
5	Marital status=2,4, Age group =1,3, Education =1,2, Gender=2	2.44(1.45)	2.35(1.37)	2.38(1.20)
	Overall	1.97(1.08)	1.93(1.08)	2.21(1.20)

[a] Category codes are as defined in Section 4.2.4

4.5 Duration of work trips

Daily commuting time is also an interesting aspect in travel behavior. Focusing on the work travel purpose, 75% of the data set was used to classify individuals into homogenous groups following the duration of travel per person (travel participant) per day. The classification scheme using Regression trees resulted into 5 groupings. The most important variable on which the data are split is gender, followed by education and the age of respondents. Marital status was not found to be important in explaining travel duration. Table 4.12 shows the mean travel durations (in minutes) per person per day and the standard deviations in parentheses. This is displayed following the derived categorization scheme firstly, for the training data set (75% of the data), the testing set (25% of the data) and finally for the simulated data, in the last column.

The results of the categorization scheme reveal that males aged between 35 and 54 years with at least College or University education travel for the longest duration per day followed by males in the same age group with junior high or high school education. It is expected that this group of people would be on top in terms of travel time. It is an active working age class and since most times, women tend to be more involved in the responsibilities of the household and taking care of children, men can in many cases afford to take on long distance jobs, which would then mostly require long travel time. Furthermore, it is observed that males with primary education are found to travel for the shortest duration. It can be observed that the travel duration of females tends to be homogeneous, irrespective of their age and education. Their travel time is also slightly above that of men in the lowest education category. This finding is still evidenced in everyday life where women tend o prefer to work close their home. All above observed trends are consistent across all data sets. The testing data set follows more or less the same trend but with a few deviations.

Simulation tools can be used to replicate individual travel to deduce real world travel situations and are ideal tools to analyze and test different transportation policy strategies in a controlled environment. The issue of reproducibility has been demonstrated here.

Table 4.12: Comparison of the duration of work trips per participant per day between the survey data and the simulated data

		Mean (Standard deviation)		
Group	Categorization Scheme[a]	Training Set	Testing Set	Simulated Data
1	Gender=2	25.52(21.33)	25.49(23.92)	25.85(23.91)
2	Gender=1, Education=1	21.76(18.05)	27.52(28.38)	23.51(23.79)
3	Gender=1, Education=2,3,4, Age group =1, 3	32.31(33.56)	41.02(61.09)	32.62(35.36)
4	Gender=1, Education=2,3, Age group=2	36.10(38.42)	34.57(51.27)	34.47(36.12)
5	Gender=1, Education=4, Age group=2	43.10(40.67)	51.99(59.89)	40.58(42.03)
	Overall	31.01(31.52)	33.53(45.56)	30.49(31.93)

[a] Category codes are as defined in Section 4.2.4

4.6 Conclusion

Survey data have been utilized here to simulate a travel data. Regarding the methodology used, a simulation approach based on Stopher et al. (2003) has been applied. This method has been developed further by proposing an alternative grouping technique and also expanding the method to incorporate an in-depth validation approach. In this application, simulation has been conducted for trip rates and travel duration. A simulation procedure has been set up by initially developing homogeneous categories of individuals following these attributes. This was done by completing regression tree runs for trips conducted and travel duration. Travel trips are initially conditioned on the purpose of travel in the first setting. In the second setting, trip rates are conditioned on travel mode. Lastly, another setting is considered where work travel duration is simulated. In an alternative simulation procedure that we propose, categories of individuals are specified a priori with respect to the available socio-demographic variables. For both utilized simulation approaches, distributions were then formulated for each of the obtained categories, from which samples were subsequently drawn to obtain a synthetic travel data set.

Observed for both methods, the findings demonstrate that the simulated data are replicative of actual travel survey data. This has been shown in different ways. Based on descriptive statistics, it was clear that the distributions are well preserved. No major significant statistical differences between the testing and simulated data results were observed when using the two approaches. However, with simulation approaches, one may not rule out the possibility of unstable results between simulations. Issues of robustness and stability concerns of the applied methodologies were investigated through setting up multiple experiments. Experiments to examine possible differences between different random training-testing splits of the actual travel data displayed less than 1% significant statistical differences in 1000 splits. Furthermore, still less than 1% significant differences were observed between simulated and testing data in 1000 splits. These percentages are very small and thus the result demonstrated stable results that can confidently be replicated based on this case case study. Both simulation methods employed here equally displayed robustness and stability of results, moreover the results from our approach were even more appealing. Notably, the newer method is also much easier to implement.

This motivating case study has further highlighted the quality of parameter estimates that can arise from simulated data. Results from the modeling approach where the Poisson regression model was fitted to both actual and simulated data were elaborated. It is particularly interesting to note that parameter estimates attained from simulated data are not generally different from those from actual data. Despite the few observed differences, overall, the parameter estimates were found to be roughly consistent in direction and magnitude, with similar estimates of standard errors. The ranges of the confidence intervals for the respective parameters were also observed to be relatively similar. Therefore, this procedure holds out considerable promise for supplementing or replacing the collection of larger and more costly samples of household travel data with simulated data. The results of this application have revealed an interesting finding that simulated travel data can be replicative of actual travel data. However, the results obtained here apply to the case study at hand and may not necessarily imply generality. Nevertheless, the result is supported in literature since similar methods have been applied in other regions and similar promise has been observed, the strength of simulated data is emphasized.

Possible further research lies in development of a joint model that handles all travel attributes simultaneously. It could be possible to set up a single model that simulates all variables of interest in a single run. This is an area that has not been explored in this current research but could potentially be useful.

The Poisson distribution that is assumed as the underlying distribution in simulation of trip rates can be viewed as being too restrictive. This is particularly because, customarily, the Poisson distribution implies that only randomness is the generating mechanism but this can be sometimes judged unreasonable in the field of transportation. Also, the Poisson imposes the restriction that the variance equals the mean and hence does not allow for large tails. In future, it would be interesting to experiment with working with another distribution with larger tails so as to account for the variation in the data more flexibly. Investigating whether statistical matching is an efficient method of extracting information if one is interested in estimation of statistical models is also a topic for further research.

5 Models for Prediction of some Travel-Related Variables

Quite often, in several studies, more data are needed than available. This challenge is encountered for the activity-based framework for Flanders; *FEATHERS*, in which richer data are required for microsimulation purposes. Synthetic populations are required to be generated for use in *FEATHERS*; Chapter 6 and 7 are dedicated to this cause. Procedures for creation of synthetic populations generally rely on aggregate-level data for the population, as well as on a micro-level sample data set. Much as aggregate data are frequently freely available, in some cases, they are lacking some variables of interest. It is for this reason, that car ownership is required to be forecast for Belgium over several years in this Chapter. More still, to lay a good foundation for the synthetic populations generation procedure, some variables need to be created within the Flemish Sample that is used as a basis for household selection. Among the variables that must be present in the synthetic population is personal income and possession of driver's license. However, these variables are non-existent in the Flemish Sample. Models are thus proposed in this Chapter, for the prediction of these variables. *FEATHERS* also requires an additional micro-level file that contains data on car-level information such as car mileage, age of car and fuel type. A model will be developed to simulate data for these variables. The work done in this Chapter is part of the *input* component of the integrated model for population synthesis proposed in Chapter 7.

This Chapter is organized as follows. Section 5.1 focuses on prediction of car ownership which is required for Belgium at an aggregate level up to the year 2021. This is followed by a description of a car mileage model in which a simulation approach is set up in Section 5.2 to simulate the necessary data at a micro-level. Models for prediction of personal income and possession of driver's license are proposed in Section 5.3 and finally, Section 5.4 concludes the Chapter.

5.1 Prediction of Car Ownership

Travel involves the use of several modes including car, slow modes such as foot and bicycle, different public transport modes, among others. Car ownership and use is important in explaining travel, since car tends to be the most popular mode of travel in many countries. It is commonly relied on to distinguish between sustainable and non-sustainable mobility. In Belgium, car mode is the most common travel mode. For instance, in Flanders, about 64.5% of the daily number of trips made by individuals is by car. The volume of vehicles has also been relatively high in the recent years. The total fleet of vehicles in the entire country increased by 47% from the year 1977 to 2006. During the latter year, there were over 6 million automobiles comprising of the vehicle fleet in Belgium (ECODATA, Federale overheidsdienst Economie and Energy, 2009). Similarly, car possession in Belgium has also followed suit with a general increasing trend over the past years. The number of households with access to a car increased from 72.4% in 1991, to 76.7% in 2001 (Belgian Federal Government, 2009b). In the Netherlands, car ownership was also high in the year 2001, and amounted to about 75.7% of the households. In comparison to some other countries in Europe, these car ownership levels are observed to be quite high. In Spain for instance, in the year 2000, the number of families with at least one car was 72.6%, with 17.7% of them owning more than one car (Matas et al., 2009). In Great Britain, the percentage of households with access to at least one car was 74% in 2001, with 28% of households availed with two or more cars in the same year (Whelan, 2007). However, Ireland displays exceptionally high car ownership levels as compared to all countries quoted above. A study by Nolan (2010) established that in Ireland, the proportion of households with one or more cars grew from 74.6% in 1995 to 80.8% in

2001. These values were based on the estimation sample they had available. A little far away from Europe, Taiwan had only 2.3 million registered cars in 1990, while in 2008 these figures had increased to 6.7 million, representing an almost tripling, over less than two decades (Chiou et al., 2009). Considering the current economic and social conditions, car possession is likely to generally increase over the future years in many countries.

In a broad perspective, increasing trends of car ownership have implications on transport policy and planning, the development and planning for land-use, fuel consumption and emission levels following from car usage (Whelan, 2007; Steg et al., 2001). These emission levels come with far reaching impact on the environment and health. There are often land-use costs and these tend to translate into high costs of driving especially in highly populated areas. More so, putting the necessary infrastructure in place requires a high monetary expenditure. To the private user, car transport involves high costs due to fuel consumption, taxes, maintenance and other expenses. On the one hand, the car mode involves several private and public benefits that stimulate its usage (Steg et al., 2001). Among these, the car provides a high flexibility in making travel choices in situations of infrastructure availability and instant demand is supported due to its availability. People also tend to have a positive attitude towards car use as a mode of travel, given its speed, independence and comfort.

The interrelationship between car ownership and car use is an important one and different forecasting models have been developed to investigate this relationship (Train and Lohrer, 1983; Jong, 1990; Chiou et al., 2009). Furthermore, there is a growing interest in research on automobile ownership and fleet volumes. Car ownership forecast is essential for many actors including transport policy planners, environmentalists, vehicle manufacturers and users. This has made it one of the most intensely researched topics in the transportation field in the past, and it continues to draw the attention of many researchers. Several models to predict car ownership levels as well as vehicle fleet have been developed by several researchers for different countries and regions in the past (Tanner, 1962; Kain and Beesley, 1965; Tanner, 1977). Also recently, this area of research continues to advance with many theories and methodologies being developed and applied.

Car ownership can be analyzed by use of either aggregate or disaggregate models. A number of research initiatives have focused on disaggregate level model development of car ownership. These models commonly rely on socio-demographic and economic characteristics of households or individuals, their location characteristics as well as availability of other modes of travel (e.g. public transport) as explanatory variables. Jong et al. (2004) give a detailed discussion of different car ownership models and different model types are compared on a number of criteria including level of aggregation, dynamic versus static model, inclusion of car use, data requirements, treatment of socio-demographic variables among others. Research (Zhao and Kockelman, 2000; Mohammadian and Miller, 2003; Potoglou, 2008) has investigated households'/individuals' automobile type choices to identify the factors influencing vehicle holding behavior. The most common socio-economic factors used in car ownership models include gender of the householder, age of the householder, household size as well as the number of children in the household (Golob and Burns, 1978; Matas and Raymond, 2008; Train, 1980). Train (1980) developed models of mode choice and auto ownership and incorporated several explanatory variables. Golob and Burns (1978) research was also cross-sectional and investigated the effects of transportation service on automobile ownership in an urban area. Giuliano and Dargay (2006) using data from the USA and Great Britain, find that population density is an important determinant of car ownership. In line with location, the size of the municipality has been seen as a proxy for a range of factors affecting car ownership (Matas and Raymond, 2008).

Given the discrete nature of the car ownership decision, discrete choice methodologies such as logistic and multinomial probit and logit are often employed in many studies (Whelan, 2007; Matas and Raymond, 2008; Bhat and Pulugurta, 1998; Potoglou and Kanaroglou, 2008) in modeling the determinants of car ownership or to explore household's decision to own cars. Some studies have been of a less static nature (Romilly et al., 2001). A recent study, Whelan (2007), utilizes a discrete choice model and individual household data on car ownership from the Family Expenditure Survey and the National Travel Survey over the period 1971-1996. In addition, a number of recent papers (Dargay, 2001; Dargay and Gately, 1999; Dargay and Vythoulkas, 1999; Nolan, 2010) have utilized repeated cross-sectional data with a view to gaining more accurate estimates of households dynamic decisions with regard to car ownership. Modeling car and motorcycle ownership may also adopt more complex discrete choice models such as the Cross-Nested Logit (CNL) or Generalized Nested Logit (Wen and Koppelman, 2001) that has a high degree of flexible correlation structure accommodating differential cross-elasticity of pairs of alternatives. However, most of these models, though interesting, they are not applicable in this study owing to the research goal and the data available.

This study aims at forecasting car ownership and vehicle fleet for Belgium up to the year 2021 at an aggregate level. These forecasts are required for creating enriched synthetic populations for Belgium that are needed in *FEATHERS*. Much as it would be interesting to develop and employ methods that make use of repeated cross-sectional data, the currently existing data do not permit carrying out such an exercise. This would have permitted incorporating a lot of information extracted through accounting for several variables while exploiting the temporal component in the data as well. Currently, for some of the variables of interest, the marginal distributions can be retrieved from national data sources. However, for car ownership, which plays an important role in the generation of a synthetic population for Flanders, these marginal distributions are not readily available and need to be predicted. Since only very limited data are currently available for Belgium on car ownership, data from the Netherlands (CBS, 2009) were used on which models were built to forecast car ownership in Belgium. Given the goal of this study and the fact that the relevant data are available on an aggregate level, an aggregate approach will be employed to solve the problem at hand. This study thus proposes an aggregate car ownership model. Interest is focused on access to car by households at aggregate levels rather than on multiple-vehicle households. Vehicle fleet levels are also forecast up to the year 2021 based on the Belgian data that are available for several years. This further adds general insight to the automobile volumes in the country. The models employed in modeling car ownership and vehicle fleet include: linear regression, Holt's linear exponential smoothing, autoregressive moving average (*ARMA*) model and Box-Tiao models.

5.1.1 Data

Data on car ownership were only available for the year 1991 and 2001 in Belgium (Belgian Federal Government, 2009b). This is a major limitation since these data are not sufficient for accurate forecasting. Data on car ownership rates are available in the Netherlands, from 1985 to 2007 (CBS, 2009). A decision, was made to borrow data from the Netherlands on which models will be built to forecast car ownership in Belgium. Much as this may not be considered a very ideal choice, it provides a plausible solution to the problem at hand. Moreover, given the proximity of Belgium to the Netherlands, several socio-demographic and economic characteristics as well travel behavior of individuals and households tend to be comparable between the two countries. To aid forecasting of car ownership patterns in Belgium in absence of the necessary data, car ownership is thus assumed to be similar between the two countries. Indeed, for the years for which data are available in Belgium, this assumption is supported. In 1991, 72.4% of the households in Belgium had access to at least one car as compared to 73.3% in the Netherlands. By the year 2001, car ownership had reached 76.7% and 75.7% for Belgium and the Netherlands respectively. Thus, data on car ownership in the Netherlands were directly inherited and used as actual data for Belgium without the need for a major updating.

Concerning automobile fleet, data are readily available for Belgium from the year 1974 to 2006 (ECODATA, Federale overheidsdienst Economie and Energy, 2009). As thus, no data borrowing was required for modeling vehicle fleet. A careful exploration of these vehicle fleet data reveals that data from the year 1974 to 1976 could not be relied on as it represented outliers which could not be supported by known facts or economic trends. These data were thus not used in the model building. More data on net taxable income of the population derived from tax forms were also existent from the year 1976 to 2003 (ECODATA, Federale overheidsdienst Economie and Energy, 2009) and were utilized in modeling vehicle fleet.

5.1.2 Methodology

The Different Models

Classical regression (Neter et al., 1996) is a well established and relatively easy methodology to implement but it is inadequate when the underlying assumptions (such as independence of the error terms) are violated, which is usually the case in time series data. In the past dynamics that tend to occur in these data such as correlation, resulted into proposition of more advanced models. Given this deficiency, several other methodologies were utilized in this study to model car ownership and vehicle fleet. Exponential smoothing and autoregressive moving average (*ARMA*) modeling techniques were initially employed to model car ownership accounting for possible autocorrelation between the observations of these series. Furthermore, a Box-Tiao model was also estimated, explaining the response series by means of input (explanatory) variables and accounting for autocorrelation among error terms. Autocorrelation is a common phenomenon in time series data, since very often, the errors may not be independent. When the errors are correlated, this implies that each error is correlated with the error immediately before it. Autocorrelation and cross correlation functions (ACF and CCF) are some of the tools that may be employed to study relations that may arise within and between time series at various lags (Shumway and Stoffer, 2005; Yaffee, 2000). The *ARMA* models have been applied before in the fields of transportation. Cools et al. (2007) applied *ARMA* exponential smoothing and Box-Tiao models to investigate the effect of holidays on daily traffic counts whereas (den Bossche et al., 2004) employed regression models with *ARMA* errors in the investigation of frequency and severity of road traffic accidents. Similar models have also been applied in modeling and forecasting vehicular traffic flow (Williams, 2001; Williams and Hoel, 2003).

Exponential Smoothing Models

Exponential smoothing methods seek to isolate trends or seasonality from irregular variation. The models incorporate these patterns into the forecast, if they are found, using weighted averages of the past data for forecasting with the effect of recent observations declining exponentially over time. In forecasting the implication of this is that recent observations are given more weight than observations further in the past. Exponential smoothing methods are noted to be good for short-term series as they are based on short-term moving averages (Yaffee, 2000).

The equation for a simple exponential smoothing model is a linear constant model with a mean and an error term. Simple exponential smoothing fails to account for trends in the data and remains incapable of handling interesting and important non-stationary processes. However, Single exponential smoothing is often better than the naive forecast based on the last observation. Nevertheless, when trends or seasonality are involved, Holt or Winters methods tend to perform better. With focus on the Holt's linear exponential smoothing model that was applied, the method is described as follows.

The Holt's linear exponential smoothing model (Yaffee, 2000) is an improvement of the simple exponential smoothing model in that, it incorporates a trend such that the most recent data are weighted more heavily than data in the early part of the series. Let Y_t denote a series at time t. The formulation of the model is then represented as

$$Y_t = \mu_t + \beta_t + \varepsilon_t \tag{5.1}$$

with the mean μ_t and the trend parameter β_t forming the two components of the equation that are smoothed and ε_t is the error term. The smoothing equation for the mean is thus

$$\mu_t = \alpha Y_t + (1 - \alpha)(\mu_{t-1} + \beta_{t-1}) \tag{5.2}$$

where the α coefficient is the smoothing weight for the equation. The smoothing equation for the parameter is given as

$$\beta_t = \tau(\mu_t - \mu_{t-1}) + (1 - \tau)\beta_{t-1} \tag{5.3}$$

where the τ is the corresponding smoothing weight. The range of the values for both smoothing weights is between 0 and 1. The rule of thumb is that smaller smoothing weights are appropriate for series with a gradually changing trend, whereas larger weights are appropriate for volatile series with a rapidly changing trend. Experience has shown that good values for the smoothing weights are between 0.05 and 0.3 (Brocklebank and Dickey, 2003).

ARMA Models

Without doubt, the *ARIMA* modeling methodology, which was popularized by Box and Jenkins (1976), is the most powerful and sophisticated methodology for forecasting univariate series (Brocklebank and Dickey, 2003). *ARIMA* models are used in time series analysis to describe stationary time series (Shumway and Stoffer, 2005; Yaffee, 2000). Like Exponential smoothing models, they attempt to explain present and future values of a series as a weighted average of its own past values. The model is a combination of an autoregressive (AR) model and a moving average (MA) model.

An AR model of order p, $AR(p)$, of a random process Y_t in discrete time t is defined by

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + w_t \tag{5.4}$$

where $\phi_1, \phi_2, \ldots, \phi_p$ are weights of the AR terms, α is a constant and w_t is a white noise series with mean zero and variance δ_w^2. Another form of the $AR(p)$ model arises from using a backshift operator B_i on Y_t, defined as $B^i(Y_t) = Y_{t-1}$. The $AR(p)$ model is then written as

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)Y_t = \alpha + w_t \tag{5.5}$$

Alternatively,

$$Y_t = \alpha + \phi_1 B Y_{t-1} + \phi_2 B^2 Y_{t-2} + \ldots + \phi_p B^p Y_{t-p} + w_t \tag{5.6}$$

The autoregressive parameters of a stationary process must reside within the bounds of stability. That is, the absolute values of the parameter estimates have to be less than unity (Yaffee, 2000).

The series Y_t can also be modeled as a moving average process in which case, the moving average model of order q, $MA(q)$, assumes that the white noise or errors w_t are combined linearly to form the observed data. The model is then defined by the following expression

$$Y_t = \alpha + w_t - \vartheta_1 w_{t-1} - \vartheta_2 w_{t-2} - \ldots - \vartheta_q w_{t-q} \tag{5.7}$$

where $\vartheta_1, \vartheta_2, , \vartheta_q$ are parameters that determine the moving average process. Alternatively, using the backshift operator as earlier defined, the $MA(q)$ process may be equivalently defined as

$$Y_t = \alpha + (1 - \vartheta_1 B - \vartheta_2 B^2 - \ldots - \vartheta_q B^q)w_t \tag{5.8}$$

The general mixed autoregressive moving average model when series Y_t are modeled as combination of an $AR(p)$ and $MA(q)$ is then termed as an $ARMA$ (p, q) process. The model is then formulated as

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)Y_t = \alpha + (1 - \vartheta_1 B - \vartheta_2 B^2 - \ldots - \vartheta_q B^q)w_t \tag{5.9}$$

The time series Y_t is assumed to be stationary and $\phi_p \neq 0$, $\vartheta_p \neq 0$, and $\delta_w^2 > 0$, where the parameters p and q are called the autoregressive and moving average orders. *ARIMA* models are, in theory, the most general class of models for forecasting a time series which can be stationarized by transformations such as differencing and logging.

In the application of this study, the *ARIMA* model was used to model and predict values in a response time series as a linear combination of its own past values and past errors. The white noise test was based on the test statistic suggested by Stoffer and Toloi (1992), which is a modification of the standard Ljung-Box test statistic. This is an approximate statistical test of the hypothesis that none of the autocorrelations of the series up to a given lag are significantly different from zero (SAS Institute Inc., 2004). Estimation of the parameter estimates of the model is effected using maximum likelihood estimation.

Box-Tiao Models

As noted before, regression models, when applied to a time series, they tend to be inadequate due to frequent violation of the independence assumption of the error terms. This violation of one of the underlying assumptions of linear regression increases the risk for erroneous model interpretation, because the true variance of the parameter estimates may be seriously underestimated (Neter et al., 1996; Cools et al., 2007).

The Box-Tiao models then become useful in solving this shortcoming that is as a result of autocorrelation. A Box-Tiao model approaches this problem by describing the error terms of the Linear regression model by an $ARMA(p, q)$ process thereby correcting for autocorrelation. In literature, several names are used to refer to Box-Tiao models. They are often called transfer function models, intervention models, interrupted time series models, regression model with *ARMA* errors, *ARIMAX* models and Pankratz (1991) refers to these models as dynamic regressions.

To define the Box-Tiao model, let

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_n X_{n,t} + \gamma_t \tag{5.10}$$

be a regression model, where

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\gamma_t = (1 - \vartheta_1 B - \vartheta_2 B^2 - \ldots - \vartheta_q B^q)\varepsilon_t \tag{5.11}$$

and ε_t are assumed to be white noise. The Box-Tiao model is then formulated as

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_n X_{n,t} + \frac{(1 - \vartheta_1 B - \vartheta_2 B^2 - \ldots - \vartheta_q B^q)}{(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)} \varepsilon_t \tag{5.12}$$

Estimation of the parameter estimates of the model is done using maximum likelihood estimation, which has been shown to yield more accurate results with similar models (Brocklebank and Dickey, 2003). Noteworthy, if several input series are used, the forecast errors for the inputs should be independent; otherwise, the standard errors and confidence limits for the response series will not be accurate. More still, when differencing of the error terms is required to achieve stationarity, all dependent and independent variables should be differenced (Cools et al., 2007; Pankratz, 1991). The requirement of stationarity applies to the noise series. If there are no input variables, the response series (after differencing and minus the mean term) and the noise series are the same. However, if there are inputs, the noise series is the residual after the effect of the inputs is removed.

Criteria for Model Validation and Evaluation

Since the main goal of this research is forecasting, validation is an important part of this process. Model validation or assessment is performed here to ascertain whether predicted values from a given model are likely to accurately predict responses on future observations or observations not used to develop the model. Different modes of validation are common in literature. These include: an external mode of validation that involves use of different sets of observations for developing the model and testing; internal validation modes such as apparent or evaluation of fit on the same data as used to develop the model and data splitting techniques plus its extensions; and resampling methods.

The data splitting (Yaffee, 2000) approach is applied here. To do this, the data available in this study were divided into 2 sets with the first part (75% of the observation) forming the training data on which the forecast model is built and the second part forming the testing/validation set (25% of the observation). These percentages are arbitrarily chosen but also common practice in validation studies (Wets et al., 2000; Moons, 2005). A base case scenario is thus run for 2003-2007 and compared with the observed car ownership patterns during this period. Similarly, validation is done for vehicle fleet.

Literature (Yaffee, 2000; SAS Institute Inc., 2004; Makridakis et al., 1997) offers several statistics for evaluating the goodness of-fit of a model. Here, to evaluate and determine which model performs better among all the applied models, 3 main criteria were used: the Akaike Information Criterion (AIC), the Mean Absolute Percentage Error ($MAPE$) and the Mean Square Error (MSE). With these statistics, different time series forecasts may be described and evaluated comparatively (Makridakis et al., 1997). The statistics are calculated as follows: The AIC

$$AIC = -2 \times log(Likelihood) + k \tag{5.13}$$

where k is the number of free parameters. Models with lower AIC values are considered to be more appropriate.

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{(y_t - \hat{y}_t)}{y_t} \times 100\right| \tag{5.14}$$

and

$$MSE = \sum_{t=1}^{T}\frac{(y_t - \hat{y}_t)^2}{(T-p)} \tag{5.15}$$

where y_t and $\hat{y}_t$ are the observed value and the predicted values respectively at time t; T is the total number of observations and p the number of parameters in the model. The above statistics are noted to be useful in comparatively evaluating a described time series forecast (Makridakis et al., 1997).

5.1.3 Results and discussion

Different models (as laid out in the methodology section) were fitted to the available data so as to provide the most accurate forecast of car ownership patterns as well as vehicle fleet in Belgium. Models are developed on the training data sets and validation is conducted based on the testing data set.

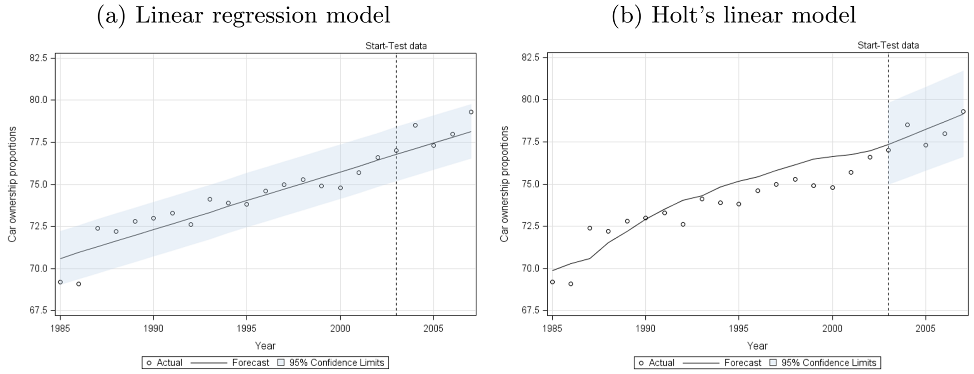
Model Application for Car Ownership

The fitted linear regression model for car ownership with a trend effect yielded an R^2 of 0.84, implying that about 84% of the variability in car ownership can be explained by time. Figure 5.1(a) shows the car ownership proportions and the corresponding forecast values with confidence intervals for the linear regression model. The forecasts do not appear to dramatically differ from the true values. However, as discussed earlier, it may not be expected that a linear regression would be the best model in predicting this series as the model ignores the existent correlation between errors. A Holt's linear exponential smoothing model may be a better alternative.

The proposed Holt's linear exponential smoothing model incorporated a linear trend with the estimated parameter for the permanent component $\hat{\mu}_1 = 79.15$ and the parameter for the linear component $\hat{\beta}_1 = 0.45$, indicating an increasing trend for car ownership. The increasing trend is nothing surprising as it is clearly depicted in the data. A 'good' model should certainly be able to at least pick up this trend. The current method uses smoothing equations for updating the parameters. These smoothing parameters for the permanent component and the linear component were fixed at $\alpha = \tau = 0.106$. The car ownership patterns are predicted, for which the effects of smoothing can be expounded on further by the graphical examination of the forecast plot shown in Figure 5.1(b). The predictions appear reasonable with the forecasts quite close to the actual data although the confidence bands are relatively wide. These confidence bands are shown for the testing data set only due to the nature of output obtained from this model. Other models considered here are implemented differently and therefore permit the bands to be shown for the training set as well.

The *ARMA* modeling was also investigated for the prediction of car ownership. In these analyses, stationarity of the series was achieved through differencing. Different analytical tools including the autocorrelation function, the partial autocorrelation function and the inverse autocorrelation function are examined to determine which *AR* and *MA* factors are required to build the *ARMA* model. The obtained final model comprised of a *MA* process of order 1 denoted as *ARMA*(0,1) or *MA*(1) for which one order of differencing was required to render the series stationary. The intercept was excluded from this model. This follows the argument that with differencing, the intercept is interpreted as a deterministic trend which is not always realistic (SAS Institute Inc., 2004). Also, when differencing is performed, the mean is often, but not always zero (Shumway and Stoffer, 2005; Brocklebank and Dickey, 2003). The parameter estimate for the *MA* factor was significant at the 10% level of significance but not at the 5% level. Nevertheless, including this term improves the predictions much more as compared to relying on simple linear regression. The white noise test was performed on the final residuals. The test statistics as suggested by Stoffer and Toloi (1992) fail to reject the no-autocorrelation hypothesis at the 5% level of significance suggesting that the residuals are white noise and that the model is adequate for this series. Figure 5.1 (c) displays the forecast plot for this model. Here, despite the fairly good forecast, the forecasts in a long term are likely to be masked with high uncertainty due to the seemingly increasing confidence band.

The Box-Tiao modeling approach offers an opportunity to handle the problem of autocorrelation as well as accounting for possible input or explanatory variables that may be influential in explaining car ownership. Differencing of the inputs is performed since when differencing of the error terms is required to obtain stationarity, all dependent and independent variables should be differenced (Cools et al., 2007; den Bossche et al., 2004; Pankratz, 1991). Innitially, the independent variables controlled for included household size, vehicle fleet and income. Incorporating other inputs such as vehicle fleet, income, did not improve the model and also entailed correlation problems between inputs. The best model thus incorporated a single household effect. The white noise test was performed on the final residuals. Also, for this model, the test statistics fail to reject the no-autocorrelation hypothesis at the 5% level of significance indicating that the residuals are white noise and thus supporting the model adequacy for this series. Table 5.1 shows the parameter estimates for car ownership using the Box-Tiao model. The results show a positive effect. The model provided predicted values that represent a substantial respectable match between forecasts and actual values for both the training and testing set of data. The visual impression from Figure 5.1(d) clearly points to the superiority of the Box-Tiao model over the previous models, with the confidence bands also much narrower than for the plots in Figure 5.1(a) to Figure 5.1(c).

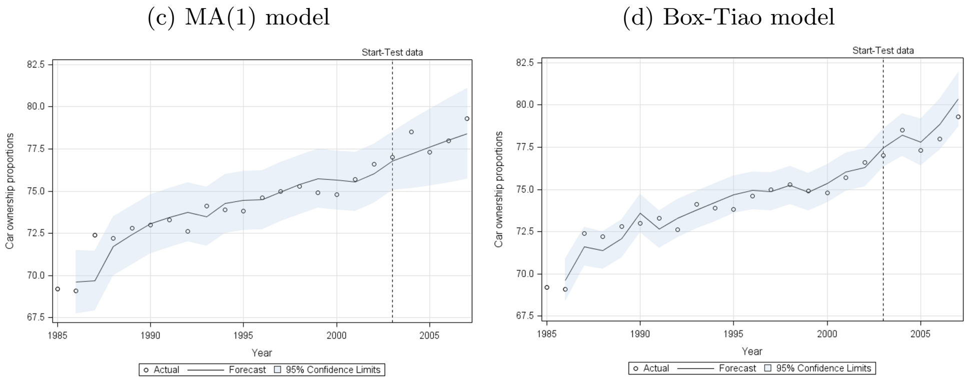

(a) Linear regression model

(b) Holt's linear model

Figure 5.1: Car ownership proportions and the corresponding forecast values with confidence intervals using different models.

Model Validation and Evaluation for Car Ownership

The different models that were built in this study were compared based on the training set of data. The testing set of data was then utilized for model validation purposes. The different criteria that were employed to assess the fit of the models are shown in Table 5.2. For the training data, according to the *AIC* criterion, the Holt's linear exponential smoothing model out-performs the other models. It is noteworthy however, that unlike the other models, the implementation of this model is based on a different way of calculating the *AIC*. Considering the results for the distance-based criteria, reveals that the Box-Tiao model is the best model, which indicates that accounting for the single household size effect adds valuable insight into the car ownership model. Cools et al. (2007) who applied similar models in investigating the effect of holidays on daily traffic counts also discovered similar performance results of the models. The validation set of data shows the superiority of the Holt's linear


(c) MA(1) model

(d) Box-Tiao model

Figure 5.1: Car ownership proportions and the corresponding forecast values with confidence intervals using different models [continued].

exponential smoothing model with the Box-Tiao model trailing very closely. Nonetheless, all the other models also provide pretty accurate forecasts as their forecast errors are all relatively small and fall within the same range. Figure 5.1 offers a visual of the different forecasts for the testing data set. Here, the Box-Tiao model reveals the narrowest confidence bands underscoring its good performance. Moreover, this model is intuitively expected to offer better results given the fact that it utilizes more information.

Model Application to Generate Car Ownership Forecast for Belgium

Given the adequate validation of the models, the Box-Tiao model is proposed proposed for forecasting of car ownership over future years. The model was thus applied to complete the task of generating forecasts of aggregate car ownership in Belgium for future years up to 2021. For comparison purposes,

Table 5.1: Parameter estimates from the Box-Tiao models for car ownership and vehicle fleet

Parameter	Estimate (s.e)	t-value	p-value
Car Ownership			
Moving average (Lag 1)	0.480 (0.238)	2.01	0.0443
Single Household	0.536 (0.078)	6.87	<.0001
Vehicle fleet			
Mean	2.601(0.150)	17.41	<0.0001
Moving average (Lag 1)	-0.681(0.201)	-3.38	0.0007
Auto regressive (Lag 1)	0.751(0.173)	4.33	<0.0001
Income (billions)	0.027(0.002)	12.27	<0.0001

Table 5.2: Criteria for model comparison

Model	Training data set			Testing data set	
	MSE_{model}	$MAPE_{model}$	AIC_{model}	$MSE_{forecast}$	$MAPE_{forecast}$
Car Ownership					
Linear Regression	0.68	0.89	45.95	1.12	0.80
Holt's linear	1.25	1.25	5.95	0.68	0.73
MA(1)	0.76	0.74	46.34	0.88	0.69
Box-Tiao	0.31	0.61	31.37	0.80	0.81
Vehicle fleet					
Box-Tiao	0.002	0.895	-59.40	0.046	2.201
Holt's linear	0.009	1.577	-92.99	0.083	3.199

also the results for the Holt's linear exponential smoothing model were shown and the graphical display of the two sets of forecasts is given in Figure 5.2. Overall, there is a forecast increase in household car ownership proportions over the forecast period in Belgium. The results from the Box-Tiao predict an increase of about 4.98% from 79.32% of households with access to at least one car in 2008 to 84.30% of households in the year 2021. Other studies in

Europe have also made similar projections. Whelan (2007) also predicted a general increase in car ownership for households in Great Britain until 2031. According to his findings, 22.3% of households in Great Britain will still have no access to car implying a 77.7% car ownership proportion in 2021. Matas and Raymond (2008) also projected a growth in the level of motorization in Spain derived from an increase in the employed population. It should be noted however that, beyond a very long term, predictions should normally be interpreted with care.

There are some possible good questions here: How far into the future will the increasing trend in car ownership go? Is there a point when this trend is likely to stop, reverse or at least slow down? If strong government intervention is effected to control this trend, for environmental or other reasons it could be possible to see a slower, or even a decreasing trend. Of course in this case, models that account for such policy interventions would be of interest. Moreover, it would be interesting to note, as De Jong and Annema (2010) asserts, policy does not always have the effect it is supposed to have, and sometimes policy implementation can take a long time. It is evident that expectations of the future are largely shaped by the present and that the current growth plays an important role in estimating future growth (De Jong and Annema, 2010). De Jong and Annema (2010) obtained some interesting findings that can shade more light on future transport trends. De Jong and Annema (2010)'s research, which was on the history of the future, established that future prognoses for traffic and transport satisfactorily fulfilled their role as indicators, and this was particularly the case for the more recent scenarios, made in the 1980s and 1990s. From the prognoses made in the past, it was apparent that caution should be exercised when assuming planned policies in the prognoses. De Jong and Annema (2010)'s work which was based on scenario building, furthermore ascertained that it is important that future scenarios are based on more than one scenario and that prognoses are allowed a broad bandwidth. It was noted that when the scenario studies for the 1980s and 1990s are compared to the year 2010, it is apparent that the results of these studies can differ from each other by 'tens of percentage points'. These 'tens of percentage points' seem to provide a good impression of the uncertainly of future estimations for periods of 10 to 20 years. Prognoses that are based on only one scenario and no bandwidth or only a relatively small bandwidth, do ignore the high degree of uncertainty in the future. De Jong and Annema (2010) further note that such findings could possibly mislead policymakers. Regarding the future trend of car ownership, Pendyala et al. (2009) studied the trends in India and the United States. The study established that a rapid rise in vehicle ownership had been experienced in India during the past decade and that the trend is likely to continue over the years to come. Whereas vehicle growth in India surpasses that in the United States by far, over the 50 year time span considered, in the United States, the automobile is the most dominant mode of transportation for nearly all travel.

While aggregate approaches, such as the one proposed here, can be suitable for projecting a continuation of current trends, they are unable to anticipate the effects of many major policy changes. The model proposed here is ideal for solving the substantial problem at hand of projecting car ownership levels but it is limited when deeper research questions are of interest. For example, it would be difficult to model the effects of introducing road pricing or to project the response to major structural changes in the economics of transportation with the current model.

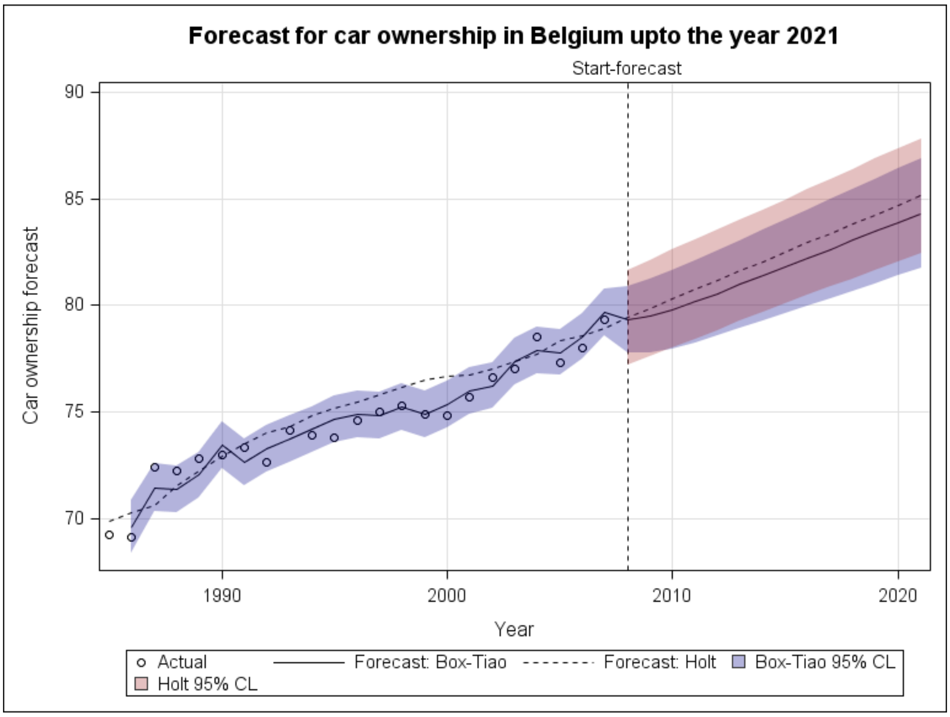
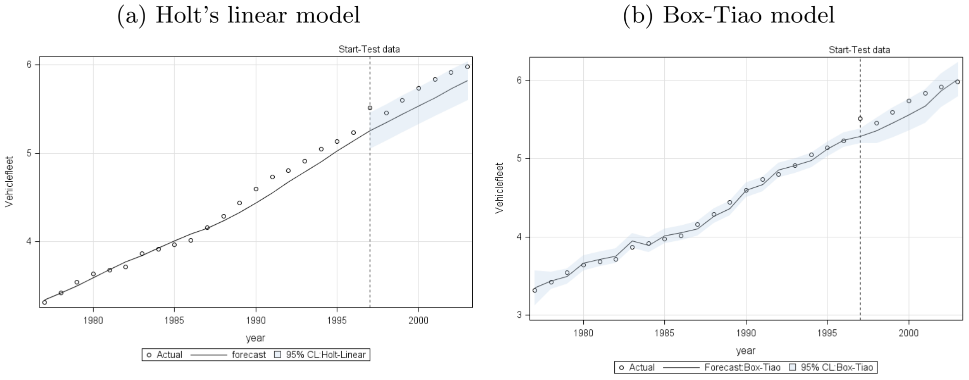


Figure 5.2: Forecast of car ownership in Belgium up to the year 2021.

Vehicle Fleet Forecast

Several models were attempted to model car fleet in Belgium. Two final models are obtained and presented here for comparison. the Holt's linear exponential smoothing model and the Box-Tiao model. The latter model clearly merged out better than the former based on both the *MSE* and the *MAPE* criteria based on both the training as well as the validation data set as shown in Table 5.2 under vehicle fleet. Based on the results displayed in Figure 5.3, it seems that the Holt's linear exponential smoothing model tends to under estimate the vehicle fleet volume over time. The Box-Tiao model on the other hand provides more accurate forecasts that fall close to the data and, more so, within a smaller confidence interval. The results from this model are shown in Table 1 under vehicle fleet. The model included an autocorrelation effect of order 1 and a moving average effect of order 1 and incorporated an income effect. No differencing was required for the series to attain stationarity. A time effect was not incorporated as it is embedded in the effect of income and incorporating the two variables results in multicolinearity problems as the two variables are highly correlated. The results show that the trend of vehicle fleet is an increasing one over time that can significantly be explained by income. Furthermore, it is established that an increase in annual taxable income in Belgium results into an increase in vehicle fleet.

The results have established an important finding that car ownership exhibits an increasing trend that will continue over the next couple of years. The model for total vehicle fleet in Belgium also offered extra support for this result as there is a direct link between this and car possession. Given the forecast at the national level, adjustments are made to obtain values at the regional level, particularly for Flanders. Car ownership levels tend to differ between regions. To be exact, Flanders displays higher levels as compared to other regions. Working within the data restrictions, a correction factor based on the available car ownership data across regions was assumed to adjust for these differences. In future, as more data become available, it would be interesting to update this correction factor, or even better still, to replace the currently used data from the Netherlands with Belgian data, if the data are sufficient for model development.



(a) Holt's linear model

(b) Box-Tiao model

Figure 5.3: Vehicle fleet proportions and the corresponding forecast values with confidence intervals using the (a) Holt's linear exponential smoothing model and the (b) Box-Tiao model.

5.2 Car Mileage model

Besides car ownership, it is of interest to generate some car-related variables at a micro-level. A simulation approach is proposed for this problem to simultaneously simulate, mileage, age and the fuel type of each car available in a household. The simulation will be based on the *FHTS* data. Within these data, a household file that specifies several variables including the number of cars possessed by a household, household size, is also linked to a car file that contains other car-related variables. On the other hand, the Flemish Sample (*FS*) has no such file for car-related data and this needs to be created. The latter sample will be the basis for sampling households in creation of synthetic data sets in Chapter 7, thus creation of these data is crucial so that car data can be available in the synthetic populations set-up.

To address the objective of creating a car file, at a micro-level, a simulation approach is set up to simulate these data. The approach has been described in detail in Chapter 4 and will not be repeated here. While in past studies that utilized a similar method, simulation of the variables of interest is performed sequentially or independently, here, in the approach that we propose, only one simulation procedure is set up to simultaneously simulate all the required variables. This is practical, since car mileage can be assumed to be a dependent variable and the other variables of interest considered as determinants. In general, understanding the determinants of household car ownership; a key determinant of household travel behavior more generally, is particularly important in the context of current policy developments which seek to encourage more sustainable methods of travel. Similarly, investigating the determinants of car mileage is important. However, the nature of the data available, often limits the extent to which the underlying behavioral influences on car mileage can be examined. This study also faces data limitations and therefore, only variables that are available can be relied on in predicting car mileage. While it would be useful to distinguish between different kinds of vehicles such as service trucks, company vehicles and private cars, data to support such an investigation are not readily available. However, this is not also the interest of the current study. Focus is therefore restricted on cars available in a household.

A *CART* model is set up to explain car mileage, with several variables including: age of car, fuel type of car, household size, number of cars in a household and age of the householder are used as explanatory variables. In so doing, only classifications for car mileage need to be developed. The simulation then follows categorizations. The simulation is done by groups based on the characteristics that arise in the final regression tree model. These characteristic groups are transferred directly to the created car file together with the simulated car mileage values to finally build a new complete car file with the corresponding variables that are involved in the model.

Car mileage is measured in kilometers as displayed on the meter. Regarding the variables used, age of car was a continuous variable measured in years. Fuel type is grouped into three groups: diesel, LPG and benzine. There was also the ‘other’ category for fuel type but it comprised of very few cars; 7 cars. It would not be interesting therefore to keep ‘other’ as a main category as it would not provide much information. Instead, we opted to re-distribute these cars to the three main types of fuel used. Many arguments can be built here on which category such cars are most likely to pass best. Based on expert opinion, it was finally decided to redistribute the 7 cars proportionally to each of the three main fuel types depending on the number of cars in each category. At the household-level, household size comprised of four categories, viz; 1, 2, 3 and 4 plus members. Number of cars is also grouped into four categories; 0, 1, 2 and 3 plus cars. The initial idea was to develop models and simulations depending on whether a car originally belonged to a household with 1, 2 or 3 cars (in the base sample) and depending on the ranking of the car in the household. However, this approach was abandoned since the ordering of cars in the *FHTS’07* does not exhibit any particular order with respect to the cars in a household. No particular distinction was made between the different cars available in a household when collecting these data. For instance, the car labeled number ‘1’ does not represent the most used, old or favorite car in the household for that matter. Therefore, modeling with respect to the first car, second car, and third car is of no relevance here since the numbering of the cars in a household have no particular meaningful order in this survey. In fact any differences that would arise would be purely by chance.

Using cross-validation, the right sized tree was decided upon after considering all the different possible sizes. Figure 5.4 shows a cross-validation plot regarding the choice made. In general, smaller and simpler trees are preferred over larger trees that have the same predictive accuracy. We therefore preferred to prune back to the smaller tree where the increase in misclassification cost is minimal. Figure 5.4 thus shows that a size of seven is realistic and optimal as it is associated with relatively low error and low complexity. The results of the categorization scheme are shown in Table 5.3 in the first and second column. An interesting and also intuitive discussion can be derived from these results. Age of car and fuel type are identified as the main variables that are important in explaining car mileage. Cars that are younger than 2.5 years display the lowest mileage with an average of just under 26,000 km. For cars that are aged between 2.5 and 5.5 years, the difference in mileage will depend on the fuel type used. Thus if a car runs on LPG or diesel, it is predicted that on average it will have 81,320 km where as if it rides on benzine, it will only have just over half of these kilometers. As cars get older, between the ages of 5.5 and 10.5 years, the difference in mileage for benzine use versus LPG or diesel, slightly narrows but LPG or diesel cars have run an average of 130,710 km as compared to benzine cars with only 90,470 km. Above 10.5 years, LPG or diesel fueled cars have over 207,000 km, which is over 1.5 times the kilometers driven by benzine cars of the same age group.

Concerning the simulation, the aim is to preserve the patterns observed in the base data. Simulations are performed by categorization group assuming the Poisson distribution. To determine the number of response values to be simulated for each group, the total number of cars possessed by households in the *FS* is multiplied by the group proportions of cars in the training set. For example, let n_{FS} be the total number of cars in the *FS* data. Let also n_i represent the number of cars in categorization group i of the training data and consequently, n the total number of cars in the training file. The number of cars in group i of the new *FS* car file, n_{FS_i} will thus be determined as

$$n_{FS_i} = n_{FS} * \frac{n_i}{n} \tag{5.16}$$

Noteworthy however, simulation is done only up to a maximum of three cars per household. Consequently, for n_{FS}, only three cars are counted for households with more than three cars. The results of the simulated data are shown in Table 5.3 in the last columns. The categorization groups of the training set are also constructed within the testing data for validation and the results are shown in column 4. Two sets of simulations are performed. For Simulated data 1, we simulate a number of units that corresponds to the total

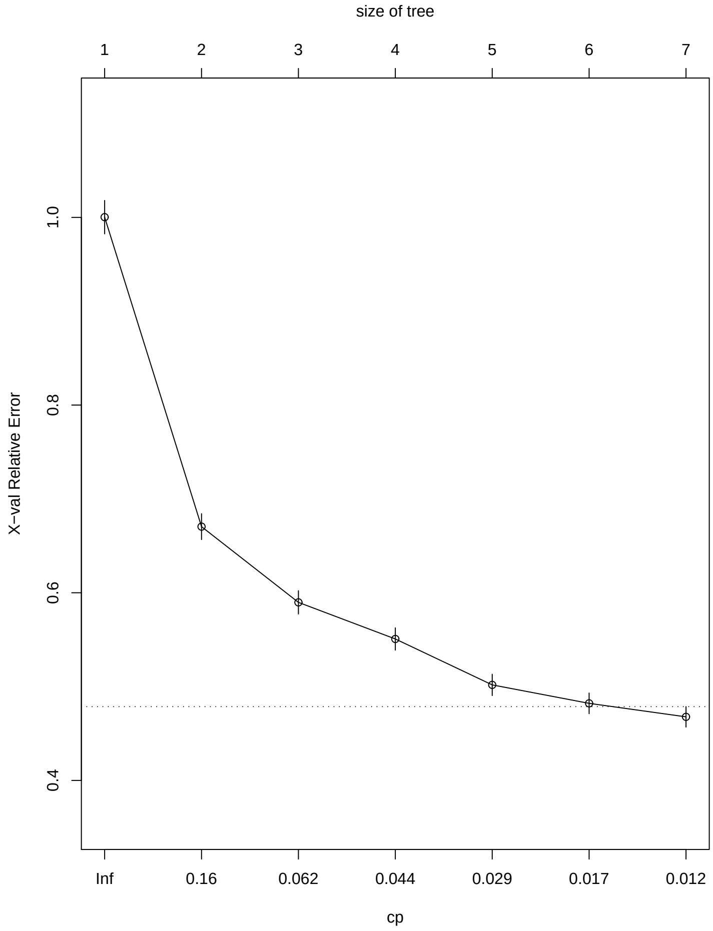


Figure 5.4: Cross validation plot.

units of a given group in the testing set. Simulated data 2 is the simulation that is made for the *FS* car file and thus the total number of units is calculated as in Equation 5.16. Thus, the created *FS* car file contains simulated data for 4294 cars. Interestingly, the simulations replicate the validation data set quite well with respect to the mean. However, for the standard deviations are notably much lower in the simulated data.

Once the simulations are done, the final step of the proposed method is to assign the simulated cars to different households. Since the simulations have been effected with the total number of cars in the sample in mind, the next challenge here is the method of assignment. Since car mileage has been simulated by group, these same groups are now formulated by making replicates of n_{FS_i} per group to correspond to the n_{FS_i} mileage simulated values of that group. Since the model did not pick up any differences in car mileage with respect to the number of cars in a household and the household size, it may be assumed that these characteristics can be ignored in assigning cars to households. The simulated data set now contains the following variables; Car identification number, age of a car, fuel type and car mileage. The observations of the simulated data are now permuted (by row) to distort the ordering of the groups. Cars are then randomly assigned to the households, in which step, the household identification numbers of the FS are linked to the synthetic car file.

The approach proposed here has been fundamental in creation of a complete car file with the variables of interest. It has provided 'nice' results that are comparable with actual data. However, the method though appealing, it is not without limitations. Unfortunately, while it would be useful to account for the hierarchy that exists within the data, the modeling approach employed here is unable to incorporate this aspect. This could be probably also the reason to why the household-level variables were not found to be important in explaining car mileage. In the analyses, use could be made of households-level variables and car-level variables. There could be correlation between cars from the same household. However, we investigated the alternative option of fitting the regression tree on different subsets of the data based on the number of cars in a household. Separating the data into sets for cars arising from households with 1 car, 2 cars and 3 or more cars, a regression tree was fit separately on each set. However, the results were not so different between the sets and therefore that alternative modeling approach was dropped. This basically suggests that the hierarchy is not so important after all, at least based on the number of cars in a household. However, there are also other interesting household variables to consider. It should be interesting to investigate use of multi-level models for this problem to see if different results would be obtained. The concern that would motivate such research would be that car units on which the data are observed can be grouped into clusters; the households, and the data from a common cluster are correlated. A potentially interesting approach to investigate in future, could be one that utilizes a linear mixed modeling framework for vehicle mileage. The method is rich in the aspect of accounting for correlation within data. We would suggest some form of sequential approach in predicting the required car-related variables. Some variables of interest may require employing other approaches such as the Generalized Linear Mixed Models. One of the main issues that may be faced here would be in considering which outcome to predict first.

Table 5.3: Car mileage (in 1000‘s of km) by categorization scheme across different data sets

		Mean (Standard deviation)			
Group	Categorization Scheme	Training Set	Testing Set	Simulated Data 1	Simulated Data 2
		(n=7974)	(n=2661)	(n=2661)	(n=4294)
1	Agecar<2.5	25.89(27.409)	27.94(28.02)	25.59(5.01)	26.534(5.22)
2	2.5 <= *Agecar* < 5.5, fuel type=‘benzine’	46.34(31.13)	43.78(26.33)	45.68(6.19)	45.36(6.45)
3	2.5 <= *Agecar* < 5.5, fuel type=‘LPG’, ‘diesel’	81.32(43.90)	83.95(44.41)	81.81(9.10)	81.38(9.24)
4	5.5 <= *Agecar* < 10.5, fuel type=‘benzine’	90.47(44.08)	88.74(47.48)	90.12(9.64)	90.17(9.83)
5	*Agecar* >= 10.5, fuel type=‘benzine’	130.71(60.01)	134.59(60.00)	130.56(11.42)	131.72(11.75)
6	5.5 <= *Agecar* < 10.5, fuel type=‘LPG’, ‘diesel’	145.64(61.95)	141.18(55.08)	145.66(11.71)	144.17(11.68)
7	*Agecar* >= 10.5, fuel type=‘LPG’, ‘diesel’	207.46(75.83)	209.61(73.57)	207.63(14.97)	207.37(14.64)
	Overall	97.99(74.37)	98.20(72.93)	97.88(55.85)	97.85(55.44)

*Significantly different from the corresponding testing data set result using the Wilcoxon rank sum test

5.3 Prediction of personal income and possession of driver's license

Given the ultimate goal of generating synthetic populations based on the *SEE'01* data file that is more complete, it was of interest to predict two main important variables: viz; possession of a driver's license and personal income earned by individuals in household. This task is important since the sample, *FS* that is used as the basis of generating synthetic data is drawn from the *SEE'01* data.

5.3.1 Personal income

A Proportional odds model is proposed here for prediction of personal income. The Cumulative Logit model for ordinal responses, the proportional odds model can be formulated as follows:

$$logit[P(Y \leq j/\mathbf{x})] = \alpha_j + \beta'\mathbf{x} \tag{5.17}$$

where $j = 1, \ldots, J-1$, the $\{\alpha_j\}$ are increasing in j, since $P(Y \leq j/\mathbf{x})$ increases in j for fixed x and the logit is an increasing function of this probability. This model simultaneously uses all cumulative logits and each cumulative logit has its own intercept. The model has the same effects β for each logit. It is argued that models with terms that reflect ordinal characteristics such as monotone trend have improved model parsimony and power (Agresti, 2002). In this study, the personal-income variable is grouped into four categories: '*zero*', '*1-1250*', '*1251-2250*' and '$>$ 2250'. The variable is thus considered ordinal and can be suitably be modeled using the proposed model. There were however issues arising at the data level that required addressing. There was a considerable amount of missing data and a multiple imputation approach was used to impute the missing data. Some assumptions are also made to improve the quality of the predictions. These include the following:

- Persons aged less than 15 years are assumed to earn no income.
- Persons who are professionally active are assumed to earn at least the lowest income.

- Persons whose work hours are non-zero are assumed to belong at least to the lowest income category.
- Persons who are aged more than 30 years are assumed to earn at least the lowest income.

These assumptions are reasonably justifiable. For instance, the under 15 would not be expected to work in Belgium. Whereas those aged more than 30 years may possibly have no job, they would still normally have some form of income from the state.

The models are trained on survey data based on which prediction are made for the population of Flanders. Given the large size of the *SEE'01* data set and resource constraints, the model is implemented by province for some larger provinces, further splits by arrondissement are conducted. Let us take an example of Limburg. After model building and model selection, the final model includes age, gender, work hours per week (*workhrs*) and work/school as explanatory variables. The variable work/school contains a value '1' if the person either works or goes to school and '0' otherwise. The variable *workhrs* simply shows the work hours worked by a person per week. The final results from modeling personal income for the province of Limburg are shown in Table 5.4. Here all variables are significant with the exception of gender. The distributions of predicted personal income are also revealed in Table 5.5 for Flanders and for Limburg as an example of the province-level distributions. It is observed that the majority (over 43%) of the Flemish people earn between 1251-2250 euros whereas the minority are in the highest income category (about 10%).

5.3.2 Driver's license

The model developed for prediction of driver's license is a logistic regression that exploits age and gender as explanatory variables. Much as more complex models were also investigated in predicting this response variable, the logistic regression model was chosen as the suitable model. Models investigated included marginal models using generalized estimating equations and Alternating Logistic Regression. These models permitted effective incorporation of household-level variables but very little or no information

Table 5.4: Parameter Estimates from the personal income model – Limburg

Effect	Parameter Estimate (s.e)	95% CI	Chi-Square	$p-$value
Intercept1	10.1628 (0.2376)	[9.6970,10.6286]	1828.83	<.0001
Intercept2	11.5431(0.2779)	[10.9985,12.0878]	1725.53	<.0001
Intercept3	14.9958(0.3844)	[14.2424,15.7492]	1521.97	<.0001
workhrs	-0.1511(0.0073)	[-0.1655,-0.1367]	423.53	<.0001
age	-0.1640(0.0072)	[-0.1782,-0.1498]	514.75	<.0001
gender	0.2509(0.1795)	[-0.1009,0.6027]	1.95	0.1621
work/school(yes)	-1.2297(0.3924)	[-1.9988,-0.4606]	9.82	0.0017

Table 5.5: Distribution of predicted personal income

Personal Income (euros)	Limburg	Flanders
zero	25.79	24.88
1250-2250	27.32	21.71
1251-2250	37.64	43.12
> 2250	9.25	10.29

was gained in comparison with use of a simpler model. Moreover, the models were highly computer intensive since their complexity was coupled with the large size of data that were used. A logistic regression was thus deemed the better option for this task.

Like ordinary regression, logistic regression extends to models with multiple explanatory variables. The logistic regression model for $\pi(\mathbf{x}) = p(Y = 1)$ at values of $\mathbf{x} = x_1, \ldots, x_p$ of p predictors is formulated as

$$logit(\pi(\mathbf{x})) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{5.18}$$

For an alternative formulation, $\pi(\mathbf{x}$ can be specified directly,

$$\pi(\mathbf{x}) = \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \tag{5.19}$$

The parameter β_i refers to the effect of x_i on the log odds that $Y = 1$, controlling the other x_j. The interpretation of the parameters is quite intuitive. For instance, $exp(\beta_i)$ is the multiplicative effect on the odds of a 1-unit increase in x_i , at fixed levels of other x_j.

It is assumed in the data, that persons aged less than 17 years have no driver's license, based on the official rules in Belgium. The results of the model revealed that both age and gender were significant in explaining possession of a driver's license. The corresponding predictions for driver‘s license were also obtained for the entire region of Flanders. The results are reflected in Chapter 7 and are not shown here.

5.4 Conclusion

The buoyant economic growth associated with the continuous construction of highway infrastructure for convenient movement of individuals and freight internationally has inevitably led to rapid growth of numbers of private vehicles during recent decades. In this study, several models have been employed and compared in modeling car ownership as well as vehicle fleet. Forecasts of the car ownership have then been made over the projection period of 2008 to 2021. These forecast were necessary for a more unique and ultimate goal of the general research of creating more enriched and accurate synthetic data sets for Flanders.

Overall, the applied models sought to isolate trends from irregular variation and to incorporate the found patterns into the forecast. The Holt's linear exponential smoothing model and the Autoregressive Moving Average ($ARMA$) modeling approaches attempted to explain current and future values of each response variable as a weighted average of the variable's own past values. A Box-Tiao model corrects for autocorrelation by describing the errors terms of the linear regression model by an $ARMA$ process. Furthermore, in addition to taking care of the past values of the response series and past errors, the Box-Tiao method offers the opportunity to model the response series using the current and past values of other series that may be useful in explaining the variability in the response. In the analyses, model comparison and validation revealed that the Box-Tiao model merged out superior as compared to the other models in modeling both car ownership and vehicle fleet providing more accurate predictions over the forecast period. The results revealed that car ownership demand will continue to increase over the next decade in Belgium.

From the year 2008 to 2021, an increase of about 5% of households with access to at least one car is predicted, with over 84% of households in possession of at least one car in the latter year. Income was found to be a strong determinant of vehicle fleet in Belgium with an increase in the annual taxable income seemingly translating into an increase in vehicle fleet.

An approach has been proposed for creation of a complete car file with the variables of interest that will be useful in *FEATHERS*, the activity-based modeling framework for Flanders. The variables that are contained in this file include car mileage, fuel type and age of a car. In the future, other variables such as car type would also be of interest to include in this file when the available data permit their creation. It should be interesting also to investigate other approaches for instance, those that would make use of multilevel models to achieve a similar objective. Such methods would account for possible correlation within clusters that is not explicitly taken care of here. The concern that would motivate such research would be that car units on which the data are observed can be grouped into clusters; the households, and the data from a common cluster are possibly correlated. A potentially interesting approach to investigate in future, could be one that utilized a linear mixed modeling framework for vehicle mileage.

In this Chapter, another goal of predicting some important variables for use in *FEATHERS* has also been pursued for the purpose of availing the lacking data for use in *Feathers*. The variables of interest were possession of a driver's license and personal income earned for which models have been built and proposed.

Owing to the general findings of this study, public transport policy has a challenge in reducing car ownership given its relationship with car use and the general vehicle fleet. High car ownership and vehicle fleet may lay a profound burden on the existent infrastructure and necessitate its expansion. The impact of increasing motorization and infrastructural expansion on the environment, health, energy demand and land-use planning may then consequently have to be considered. The trend towards greater ownership of private vehicles has not only created ubiquitous congestion on urban roadways and intercity highways, but also excessive emissions and energy consumption. Towards sustainable transportation, it is crucial to propose countermeasures capable of effectively curtailing ownership and usage of high-emissions and low fuel efficiency cars and motorcycles. The policies would thus encourage use and production of energy efficient and environmental friendly vehicles. The results of this study therefore not only benefit transport policy markers but also vehicle users and manufacturers as well as environment and health planners.

There is a notion that the demand for cars can become saturated. In future, it may be of interest to develop a model of the relationship between economic development and per capita private car ownership. Such research may also focus on investigating the factors that influence the evolution of vehicle stocks. Furthermore, if the goal is devising effective management strategies to relieve dependency on private vehicles, *i.e.* cars and motorcycles, then disaggregate choice models regarding the ownership, type and usage of cars and motorcycles may be required to achieve this. It would be interesting to model choice behaviors related not only to ownership and usage, but also to car type, since gas mileage and emission coefficient of cars differ considerably with engine size and age. It could also be of interest to investigate if energy consumption and emissions vary markedly across different engine sizes and ages.

While this research has focused on the car mode share, in the broader context, it could be interesting to investigate how other modes of transport come into play, predicting the contribution of each mode in whole network of transportation. Therefore in the context of activity-based models, future research may focus on predicting a complete vector that contains all mode shares of transport.

The model proposed in this study is ideal for solving the substantial problem at hand of projecting car ownership levels but it is limited when deeper research questions are of interest. More so, when relevant disaggregate-level data are available some other interesting models have been used in literature. Modeling car and motorcycle ownership may adopt more complex discrete choice models such as the Cross-Nested Logit (CNL) or Generalized Nested Logit (GNL)(Wen and Koppelman, 2001) that has a high degree of flexible correlation structure accommodating differential cross-elasticity of pairs of alternatives. However, the CNL or GNL model requires estimating a large number of inclusive value and allocation parameters, which may result in computational burden. When the number of alternatives is large, difficulty in estimation can be extremely severe. Chiou et al. (2009) used Multinomial Logit (MNL) and Nested Logit (NL) in their study on integrated modeling of car/motorcycle ownership, type and usage for estimating energy consumption and emissions. They argue that the NL model can be feasibly estimated with a large number of alternatives and well serve for the purpose of policy simulations. Other advanced discrete choice models such as mixed logit models (Brownstone and Train, 1998; McFadden and Train, 2000) may also be interesting options for further research.

6 The Framework for Synthetic Populations

6.1 Introduction

Activity-based travel analysis has seen considerable progress in the past couple of decades leading to the development of several comprehensive activity-based travel models. Within the field of activity-based analysis and the transportation research in general, there has been further growing interest in the generation and use of synthetic populations over the years. The development of modeling systems for activity-based travel demand ushers in a new era in transportation demand forecasting and planning. Several disaggregate land use and activity-based travel models, which represent decisions and actions of individual persons and households, incorporate microsimulations. In these models, synthetic populations are initially created and the prediction of the outcomes for each unit of the population is made. The results are then aggregated to guide policy related analysis and decision making. Microsimulation is thus increasingly playing a major role in the development of demand-modeling practice and also drawing more and more attention from the extensive transportation policy and planning community. Henson et al. (2009) provide a list of different activity-based models identifying also the particular models that incorporate microsimulation. The activity-based platforms that incorporate microsimulation include ALBATROSS (Arentze et al., 2000), AMOS (Kitamura et al., 1996), CEMDAP (Bhat et al., 2004), DEMOS (Sundararajan and Goulias, 2003), FAMOS (Pendyala et al., 2005), our model FEATHERS (Bellemans et al., 2010), Jakarta (Yagi and Mohammadian, 2008), MORCPC (Vovsha et al., 2003), RAMBLAS (Veldhuisen et al., 2000), San Francisco (Jonnalagadda et al., 2001), TASHA (Miller and Roorda, 2003) and TRANSIMS (Rilett, 2001; Nagel et al., 2003) among others.

Synthetic populations do not constitute actual data but are rather built from information or data about groups of units obtained from various existing sources and also from data on samples of the units of interest. Therefore, roughly speaking, the process of creating a 'synthetic' individual involves borrowing profiles or distributions of real individuals and fashioning or cloning another profile/distribution that is sufficiently similar. It can be argued that if a synthetic population can be generated with the same, or at least very similar, properties as the original data, the synthetic population can replace the original one in simulations and other forms of manipulation. Synthetic populations are not limited to the area of transportation research but also to a broader range of applications. There is a keen interest in using synthetic populations in many areas of medicine, epidemic research, mathematics, statistics, sociology, computer sciences and the field of economics in general (Boman and Holm, 2004; Brouwers, 2005; Boman and Johansson, 2007). Synthesis models are used to develop models for predicting, preventing, and handling extensive and serious problems such as infectious diseases and catastrophes. With synthetic populations, research can study for instance, the the spread of contagions in society (Bowman and Rousseau, 2006). This is essential as it can not be studied in reality by releasing a virus and observing how it spreads. Instead, the work necessitates a study environment, using a synthetic population generated from a real one.

A population synthesizer should be able to synthesize the required population as accurately and precisely as possible, for as many variables as possible that are known to influence travel behavior. Whenever possible, this ought to be done at geographical levels that are as disaggregate as possible. In synthesizing a population, there are often limitations to the number of inputs. Normally, at most only part of the household and individual characteristics accurately represent the population at the regional level of geographic aggregation. Thus a number of other characteristics tend to be inaccurate and imprecise for the much smaller geographic regions that were not originally in focus. Nevertheless, a population synthesizer constitutes a powerful tool and provides data that would otherwise not be available. Producing data sets of artificial units by simulation has been argued to guarantee protection, and it allows to focus primarily on data quality (Franconi and Polettini, 2007). Statistical data analysis depends primarily on parameters (*i.e.* aggregates) rather than individuals (see Little, 1993) and therefore information at the individual level can be modified with no harm provided these aggregates are maintained (Franconi and Polettini, 2007).

Creation of synthetic populations is a relatively fresh field of research implying it encompasses new methods, that are still developing and thus are yet to be fully established. At present, creating synthetic populations is a well-understood problem when units are defined on a single level, say, households or persons. The majority of the currently developed population synthesizers control for variables only at one level, usually the household level. The first synthesis procedure to be developed appears to been originally proposed by (Wilson and Pownall, 1976). Wilson and Pownall (1976) proposed a procedure in which the marginal and two-way aggregate distributions for a given zone are used sequentially to synthesize the specific attribute values for a given set of households or individuals. This procedure is straightforward and easy to implement. However, it can be problematic as it ignores the potentially significant correlations among the variables (i.e., dimensions) because only two-way correlations can be taken into account in the conditional probability structure (Miller, 2003). This is a potentially serious problem. Beckman et al. (1996) however, directly addresses this issue and deals with the problem of generating synthetic baseline populations based on sample and census data that are available in the United States. The study they conduct is adapted to the data needs of the TRANSIMS model and entails combining disaggregate data from a given source with aggregate data from another source. This approach has been integrated in most activity-based travel models (Bowman and Rousseau, 2006; Hensher et al., 2004; Hunt et al., 2004; Consult, 2005) that utilize simulation systems. It has also been generally adopted worldwide in activity-based, land use and tour-based model endeavors.

Most of the research in the field of transportation that was performed after that of Beckman et al. (1996) on synthetic populations, refers to the study as a traditional method. The study was thus a milestone for research on synthetic populations. Several disaggregate land use and activity-based travel models which represent decisions and actions of individual units have been developed for different regions. They have been developed for specific geographical areas based on some particular pre-specified combination of variables. The population synthesizers that are normally incorporated within models such as TRESIS in Hensher et al. (2004), MORPC in Vovsha et al. (2003); Consult (2005), ALBATROSS in Arentze and Timmermans (2005), TRANSIMS in Beckman et al. (1996) and LANL (2003) are designed to estimate the base year joint distribution as well as the forecast year joint distribution. Different population synthesizers vary substantially in how they estimate the base year joint distribution and with respect to the control variables used. The most common household-level control variables include household income, household size, number of workers, age of the householder and presence of children (Arentze and Timmermans, 2005; Beckman et al., 1996; Hensher et al., 2004; Ton and Hensher, 2003). Some population synthesizers (Bowman and Rousseau, 2006; Consult, 2005) use similar techniques for the estimation of both the base and forecast year joint distribution. An example of such techniques is Iterative Proportional Fitting (*IPF*). Others such as Oregon2 in Hunt et al. (2004) also apply models to simulate the evolution of the base year population. Likewise, TRESIS in Hensher et al. (2004) utilizes models to estimate changes in residential and job location, without adjusting population demographics.

Miller (2003) provides a review of the main steps of the process of population synthesis. Mohammadian et al. (2010) emphasizes that population synthesizing is to generate a list of synthetic population units (i.e., households or individuals) that are statistically consistent with the available aggregate data and moreover, all synthesis procedures developed so far draw a realization of the disaggregate population from the aggregate data Miller (2003). In a broad range of population synthesis studies, the main challenge that is frequently encountered is that of simultaneously controlling for both household and individual-level distributions. When a population needs to be synthesized on both household and person level at the same time, the problem is less well understood and the methods to handle it have thus been lacking for a long time. Recently however, several researchers have expressed interest in the topic and a number of new approaches have been proposed in literature to deal with this problem. Recently, Guo and Bhat (2007) proposed a new population synthesis procedure that generally addresses the main problems faced with Beckman et al. (1996)'s approach. The approach provides a solution that deals with simultaneous control of both household and individual-level distributions in estimating target joint distributions as well as dealing with the problem related to zero-cell values. Ye et al. (2009) further presented a heuristic approach called the Iterative Proportional Updating (*IPU*) algorithm for generating synthetic populations in a computationally efficient way. The algorithm generates synthetic populations based on a procedure whereby both household-level and person-level characteristics of interest can be matched simultaneously. In the same quest, Pritchard and Miller (2009) introduce a new method that implements *IPF* with a sparse list-based data structure that permits use of several more attributes per unit as compared to the number of attributes commonly allowed in typical applications of *IPF*. The authors further use a new approach to synthesize the relationships between units so as to form household and family units in addition to individual person units. Very recently, Auld and Mohammadian (2009) proposed a methodology for generating synthetic populations with multiple control levels. Auld and Mohammadian (2009)'s methodology determined how both household and person-level characteristics could jointly be used as controls when synthesizing populations, as well as how other multiple level synthetic populations, such as firm/employee, household/vehicle, e.t.c. could be estimated. Mohammadian et al. (2010) also presents a methodology for simulating disaggregate and synthetic household travel survey data by examining the feasibility of the spatial transferability of travel data.

Several methods have thus been proposed and more are still being coming up. However, currently, there is lack of comparative studies in which the predictive performance of the competing procedures of population synthesis is compared in a single study. Consequently, as the world of new techniques unfolds, the discussion about future directions, pros and cons of such emerging competing technologies, remains almost philosophical in nature from a predictive point of view. This thesis is partly motivated by this cause motivating a comparative review of these methods.

The major goal in this Chapter is to lay down the framework for creating synthetic populations. Several methodologies are potentially interesting to consider, however, the focus here is kept on three different approaches that will be investigated for the ultimate purpose of generating synthetic populations. Population synthesis being a new field of research, most of the methods are fresh proposals that can benefit from a critical and evaluative eye. The already existing population synthesis methods are currently not well understood and fully developed. In this Chapter, we aim at achieving the following objectives. It is the aim here, to provide an ample description of the techniques under consideration so as to add insight to their understanding. Formal mathematical formulations of the methods, which have been lacking, will be proposed following Nakamya et al. (2009). Furthermore, a conceptual comparison will be given providing more enlightenment on the methods. Therefore, contribution is expected to be made on the general in-depth and better understanding and working of the methods.

In the next Section, the data requirements for the procedures for generating synthetic populations are presented. The overall framework is then presented next, followed by a description of each population synthesizing method considered. The first population synthesis method presented follows that, which was proposed by Beckman et al. (1996). The second is a method developed by Guo and Bhat (2007) and lastly, the *IPU* procedure proposed by Ye et al. (2009). The methods are employed in the generation of synthetic populations for the Flemish region of Belgium for the year 2001, 2007 and 2021 in Chapter 7.

6.2 Data Requirements

The main data requirements for most population synthesis procedures are mainly from two sources:

- Data representing the joint distribution/multiway table for the target distribution. These joint distribution data are however, not always directly available for the population of interest. In some cases, only marginal level data may be available plus some joint distribution data from other sources, in which case these sources are used in conjunction.
- A sample data set containing disaggregate data. This micro-level data set includes all control variables of interest plus some other variables investigated in the sample survey.

6.3 The General Procedure for Generation of Synthetic Populations

Most population synthesizing methodologies start with the estimation of the target joint distribution and advance with a procedure to select units for a synthetic population. These steps will be further elaborated upon in the next sub-sections and Figure 6.1 gives a graphical display of steps involved in the procedure. Socio-demographic variables of interest in a given study are identified and controlled for at the household and/or the person-level. These are generally variables that are known to be of significant influence to the individual behavioral outcome of interest and more importantly, data need to be available for these variables. As shown in Figure 6.1 the available summary level data are used to determine the target joint distribution. That is to say, the marginal distribution of the target population is used to update an available informative joint distribution. The available joint distribution may be for instance older distributions for the same target population. In any case, this step of estimating the target joint distribution may be skipped if the target joint distribution is known. The next step in generating synthetic populations then involves drawing units such as households based on some predetermined criteria. When the desired number of population units is reached in accordance with the target joint distribution, the population synthesis procedure is then completed. Very recently, Mohammadian et al. (2010) presented a process of developing synthetic populations and simulating household travel survey data for areas where actual travel survey data are not available. Mohammadian et al. (2010)'s model involved steps similar to the general procedure stipulated here and in addition, incorporated transferability models and a Bayesian updating module to facilitate simulating disaggregate household travel data for local areas.

In this study, three main population synthesizing procedures have been considered in detail. The procedures have also been implemented using the *R* software and applied on the Flemish region of Belgium. The first algorithm to be implemented follows Beckman et al. (1996)'s procedure, followed by the algorithm by Guo and Bhat (2007) and lastly the *IPU* procedure by Ye et al. (2009). Before focusing on these procedure, the *IPF*, a general algorithm, which is applicable in many population synthesis procedures is presented first.

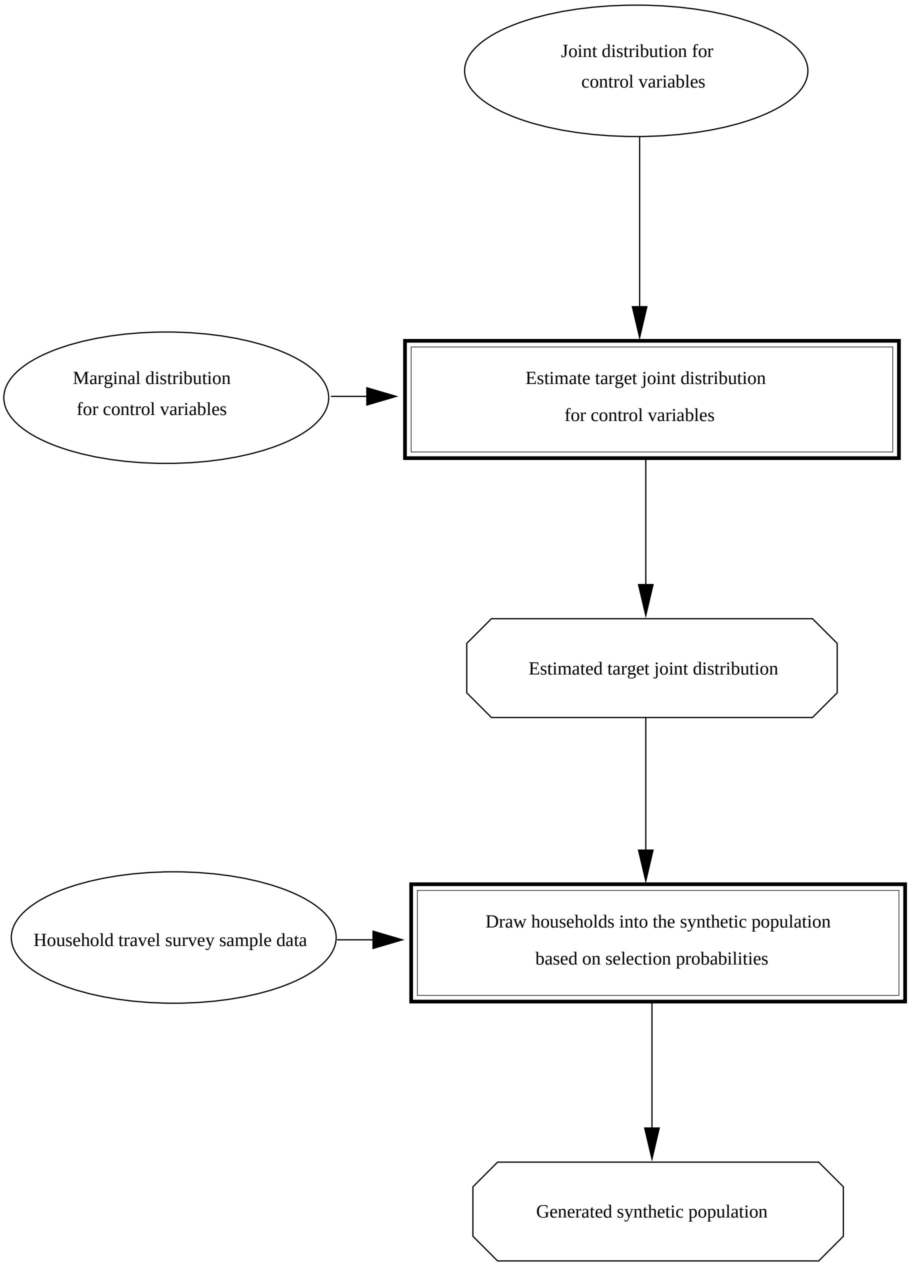

Joint distribution for
control variables
Marginal distribution
for control variables
Estimate target joint distribution
for control variables
Estimated target joint distribution
Household travel survey sample data
Draw households into the synthetic population
based on selection probabilities
Generated synthetic population

Figure 6.1: General overview of a population synthesis procedure.

6.3.1 Generation of the target joint distribution

The principal tool for estimating a joint distribution is the Iterative Proportional Fitting (*IPF*) method (Deming and Stephan, 1940). *IPF* is a well-established method and is commonly used in practice for estimating the joint distribution of a set of control variables of interest. The conventional *IPF* is used for maximum likelihood estimation in hierarchical log linear models and is also very frequently applied in transportation modeling. Originally proposed by Deming and Stephan (1940), the iterative procedure has seen many further modifications (Fienberg, 1970, 1977; Ireland and Kullback, 1968; Pritchard and Miller, 2009), citations, explorations and applications (Arentze et al., 2007; Beckman et al., 1996; Birkin and Clarke, 1988; Bishop et al., 1975; Guo and Bhat, 2007; Little and Wu, 1991; Nakamya et al., 2007; Wong, 1992). Deming and Stephan (1940) first used it to adjust frequency tables of a sample that was the source of disaggregate data, to match a known marginal distribution. Research by Fienberg (1970, 1977) has widely explored and reported the mathematical procedures involved in *IPF*. Wong (1992) reviewed and evaluated the performance of the procedure using it to generate disaggregated spatial data from aggregated data. Little and Wu (1991) report the technique's (nice) properties and denote that *IPF* has been shown to result in maximum likelihood estimates of proportions. Wong (1992) further showed that *IPF* yields maximum entropy estimates of the joint distribution under the constraints of the given marginal distributions. Birkin and Clarke (1988) provide a census-based application for the use of *IPF* in geographical research and modeling. *IPF* has also been widely used as part of the microsimulation (Clarke, 1996; Williamson et al., 1998; Williamson and Clarke, 1996) methodologies for the simulation of household attributes.

The general formulation of the *IPF* algorithm is described here using the following notation. Let us assume presence of two data files, one for the joint distribution from a sample and another for the known marginal distribution. Let the number of counts of observations from the n-way joint distribution table be represented by $C_{i_1.i_2 \ldots i_n}$ for cell $i_1.i_2 \ldots i_n$, where $i_j = 1, 2, \ldots, m_j$ is the observed value of the $j-th$ demographic variable with m_j categories. Let $Q_k^{(j)}$ denote the known marginal totals of category k of the $j-th$ demographic variable. Since the basic operation of *IPF* entails the adjustment of a count (or proportion) in a given cell, let $C^r_{i_1.i_2 \ldots i_n}$ denote the estimated count in cell $i_1.i_2 \ldots i_n$ at iteration r of the *IPF* procedure. At the beginning of the *IPF* procedure, let

$$C^{(0)}_{i_1.i_2 \ldots i_n} = C_{i_1.i_2 \ldots i_n} \tag{6.1}$$

Also, for a fixed category k of the $j-th$ demographic variable, let

$$C^{(r)}_{...i_j=k...} = \sum_{i_1=1}^{m_1} \cdots \sum_{i_n=1}^{m_n} c^{(r)}_{i_1.i_2...i_j=k\cdots i_n} \tag{6.2}$$

where the summation in Equation 6.2 is not performed for the index corresponding to the fixed category k. Equation 6.1 implies that the cell values of the joint distribution of the sample file are considered as the initial values to be updated. The *IPF* estimated cell counts $C^r_{i_1.i_2...i_n}$ at iteration r are then obtained through a procedure of updating them for each margin in turn. Thus for all values of i_1, i_2, *etc.*, and the $k-th$ category of the $j-th$ marginal, updates are derived through

$$C^{(new)}_{i_1.i_2...i_j=k...i_n} = C^{(old)}_{i_1.i_2...i_j=k...i_n} \times (Q^{(j)}_k / C^{(old)}_{...i_j=k...}) \tag{6.3}$$

where at each iteration $C^{(old)}$ is set to $C^{(new)}$ estimated from the previous iteration and the iterations continue until convergence. In Equation 6.3, the term in brackets *i.e.* $(Q^{(j)}_k / C^{(old)}_{...i_j=k...})$ can be viewed as an adjustment factor for category k of demographic variable j. At each iteration, the adjustment factor is thus recalculated for each category of a given variable in turn. Conditional on the current category of the variable under consideration, each corresponding cell value of the update joint distribution is adjusted by this adjustment factor. Convergence is achieved when the relative change difference in $C^r_{i_1.i_2...i_n}$ between consecutive iterations is smaller than a predefined small value, judged based on the desired level of accuracy. Alternatively, convergence can be said to be achieved when all adjustment factors are equal or close to 1.0. At the end of the procedure, the estimated target joint distribution is obtained. The mathematics concerning *IPF* is discussed and reported in greater depth by Birkin and Clarke (1988), Bishop et al. (1975) and Fienberg (1970, 1977).

A common challenge that arises in applying the *IPF* method is the zero-cell value problem. It may arise especially for small geographies and also when several variables or variables with many categories are used to construct the joint distribution. When this problem exists, the *IPF* procedure may fail to converge to a solution. Whereas Beckman et al. (1996) has recommended to add an arbitrarily small value to such cells so as to allow the convergence of the *IPF* procedure, Guo and Bhat (2007) argues that this may introduce an arbitrary bias. Ye et al. (2009) proposed an approach to deal with this problem that involves borrowing the prior information (probabilities) for the zero-cells from the population data corresponding to the entire region.

6.3.2 Generating the households

Procedure 1 (P1)

Beckman et al. (1996) proposed a methodology for creating a synthetic baseline population of individuals and households which can be applied in activity-based transportation models. The procedure of Beckman et al. (1996), which controlled only for household-related variables, is currently considered the conventional approach (Guo and Bhat, 2007; Pritchard and Miller, 2009) for synthesizing base year populations. This procedure basically entails two main steps in generating a synthetic population: Estimating the cross-classified table or the joint distribution and selecting sample households to construct the synthetic population of households. The basic procedure is generally based on a sample data set and summary data from the census. Based on the required socio-demographic variables, the multi-way table is constructed from the sample data set. In conjunction with the marginal distribution of the population of interest, the multi-way table is then estimated maintaining the correlation structure in the constructed multi-way table from the sample data. In the final step of the procedure, entire households are selected from the sample. These are selected with respect to the estimated probabilities of the estimated joint distribution to form the synthetic population. The tool used for estimating the joint distribution of the selected household characteristics in their procedure is *IPF*. The drawback of Beckman et al. (1996)'s approach is mainly the control for distributions at only one level. Arentze et al. (2007) developed a two-step procedure based on *IPF* for generating synthetic populations. This method uses the concept of relation matrices to convert distributions of individuals to distributions of households in a preprocessing step. Consequently, in the initial step, known marginal distributions of individuals are converted to marginal distributions of households using a relation matrix based on the relevant attributes. In this way, the marginal distributions can be controlled at the person-level as well. Based on the data that are available in a study, a relation matrix allocates persons to household attributes in a way that is consistent with person and household data. In the second step, the obtained marginal household distributions are utilized as constraints of a multi-way table of household counts. In both steps, an *IPF* procedure is applied to estimate the joint distribution and the same sample of households is used. Arentze et al. (2007) eventually synthesize the population at one level, the household. Guo and Bhat (2007) also proposed a procedure controlling for both household and person-level distributions. This method is presented in next.

Procedure 2 (P2)

Guo and Bhat (2007) proposed a procedure that deals with the limitations that are faced with the conventional approach of population synthesis of Beckman et al. (1996). The new population synthesizing procedure is characterized with generic data structures and the relevant functions that aid in overcoming the zero-cell value problem. The procedure proposed by Guo and Bhat (2007) provides an overall method that is modified from Beckman et al. (1996)'s basic algorithm to permit simultaneous control for both household and individual-level variables. The algorithm is implemented into an operational software system and used to generate synthetic populations for the Dallas-Fort Worth area in Texas. Guo and Bhat (2007)'s procedure entails estimation of the target joint distributions in the first steps followed by a set of steps for selection of households for a synthetic population. As in similar past research (Arentze et al., 2007; Beckman et al., 1996), the method used for estimation of the joint distributions is *IPF*. At the stage of *IPF*, the household-level joint distribution and the individual-level joint distribution are determined based on some control variables of interest. These joint distributions then represent the target population sizes. Guo and Bhat (2007) offer a detailed elaboration of their procedure in their paper. However, the description lacks a clear mathematical outlay that would be essential in formalization of the procedure and in aiding researchers interested in implementing the algorithm in different software applications.

Here, Guo and Bhat (2007)'s procedure of household selection is described and a formal mathematical formulation is put forward following Nakamya et al. (2010). Let K_h and K_p denote the total number of cells of the contingency tables which represent the household-level and person-level joint distributions, respectively for the control characteristics/demographics of interest. Also, let h_i and p_j, where $i = 1, \ldots, K_h$ household-type and $j = 1, \ldots, K_p$ person-type, denote the number of households belonging to group i and the number of persons belonging to group j, respectively. Groups i and j refer to the groupings of the joint distributions. For instance, for a two-way table of age and gender, a particular group j may refer to females of ages $25 - 39$ years. In a similar manner, h_i^{new} and p_j^{new} correspond to the number of households and number of persons of the synthetic population respectively. Thus, initially $h_i^{new} = p_j^{new} = 0$, for all i and j; these elements will be updated as households and individuals are drawn during the simulation process.

Apart from the joint distributions of households and persons, we also have the household and person sample data sets with corresponding sizes N_h and N_p. These data sets are linked since each household in the household-level sample corresponds to one or more persons in the person-level sample. So, the element $H_l^{(i)}$, with $l = 1, \ldots, N_h$, is the $l-th$ household belonging to a specific characteristic group i. As mentioned earlier, each household contains one or more individuals, therefore for household $H_l^{(i)}$ there are n_l persons belonging to that household in the person sample. Thus, for each $H_l^{(i)}$ there are $P_{l,1}^{j_1}, P_{l,2}^{j_2}, \ldots, P_{l,n_l}^{j_{n_l}}$ individuals each belonging to some specific characteristic group j. Following these definitions, the algorithm then proceeds as follows:

1. Compute the selection probabilities for each household $H_l^{(i)}$ using

$$Pr_l^{(i)} = \left(w_l / \sum_{s=l}^{N_h} w_s a_s \right) \cdot \frac{(h_i - h_i^{new})}{\sum_{k=l}^{K_h} \left(h_k - h_k^{new} \right)} \tag{6.4}$$

where w_l is the associated weight from the household sample data set and

$$a_s = \begin{cases} 1 & \text{if the } s-th \text{ household belongs to group } i; \\ 0 & \text{otherwise.} \end{cases}$$

2. Based on the probabilities $Pr_l^{(i)}$ randomly draw one household; $H_{l^\star}^{(i^\star)}$.

3. Retrieve the person(s) belonging to household $H_{l^\star}^{(i^\star)}$; these are $P_{l^\star,1}^{j_1^\star}, P_{l^\star,2}^{j_2^\star}, \ldots, P_{l^\star,n_{l^\star}}^{j_{n_{l^\star}}^\star}$.

4. If $h_{i_\star}^{new} < h_{i_\star}$ and $p_{j^\star}^{new} < p_{j^\star}$ for all the possible groups $j^\star$, then

   (a) add household $H_{l^\star}^{(i^\star)}$ and the constituent persons $P_{l^\star,1}^{j_1^\star}, P_{l^\star,2}^{j_2^\star}, \ldots, P_{l^\star,n_{l^\star}}^{j^\star_{n_{l^\star}}}$. to the synthetic population,

   (b) update $h_{i_\star}^{new} = h_{i_\star}^{new} + 1$, and $p_{j^\star}^{new} = p_{j^\star}^{new} + 1$ for all groups $j^\star$;

   Otherwise, discard household $H_{l^\star}^{(i^\star)}$ and the person(s) $P_{l^\star,1}^{j_1^\star}, P_{l^\star,2}^{j_2^\star}, \ldots, P_{l^\star,n_{l^\star}}^{j^\star_{n_{l^\star}}}$.

5. Repeat steps 1-4 until the synthetic population reaches the targeted size.

In practice, reaching the exact targeted size is not easy; one can use restrictions, controlling for the total sums of household groups and/or person groups, such as $\sum_{i=1}^{K_h} h_i^{new} = \sum_{i=1}^{K_h} h_i$ and/or $\sum_{j=1}^{K_p} p_j^{new} = \sum_{j=1}^{K_p} p_j$, but this restrictions are often difficult to control simultaneously. In addition, the formula for calculating selection probabilities in step 1 of the algorithm eventually reduces the selection probability of a sample household as further households of the same group are added to the synthetic population. This is an additional problem in satisfying exactly the aforementioned restrictions, since more and more household selection probabilities tend to zero as the total target sizes are approached for given characteristic types. Therefore, often an allowance is accepted for a Percentage Deviation from the target population Size (*PDS*), so that the synthetic population size is as close as possible to the targeted population size. In this case, a divergence from the target sizes is permitted in the procedure depending on the defined *PDS* level.

Procedure 3 (P3)

Ye et al. (2009) also proposed a procedure called the Iterative Proportional Updating (*IPU*) for generating synthetic populations whereby both the household and person-level attribute distributions are matched against known marginal distributions. While the preceding method relies on recalculation of household selection probabilities and evaluation of household and person-level distributions at each household draw, the current approach takes a different approach in order to simultaneously control for joint distributions at both levels. The current procedure is centered on calibration of weights based on household and person-level distributions. As in the previous method, Ye et al. (2009)'s approach involves initially estimating the target household and person-level joint distribution using *IPF*. The household weights are then adjusted based on person weights obtained from the *IPF* procedure. The authors offered the geometric interpretation of their algorithm and further laid out its general formulation. This approach was well described and some mathematical issues were addressed but the methods still required putting together in one workable mathematical framework that can easily be followed and formalized. Towards this goal, a formal mathematical formulation of Ye et al. (2009)'s procedure is offered in this thesis as follows.

Following closely the notation used earlier in the previous approach, let $\mathbf{X}_{N_h*M}$ be a special frequency matrix where N_h is the total number of households in the sample file. $M = K_h + K_p$, the total number of cells or characteristic/demographic (household and person) types. As before, K_h and K_p denote the number of demographic types tables which represent the household-level and person-level joint distributions respectively. A matrix element $x_{l,r}$ in $\mathbf{X}$ thus denotes the frequency count corresponding to the $l-th$ household for characteristic r, with $r = 1, 2, \ldots, M$. Also, let C_q be a vector containing estimates of the joint distributions (both household and person-level) such that

$$C_q = h_i + p_j,$$

where h_i and p_j, are defined as before. If W_l is a vector of household weights with initial elements $w_l = 1$, assume then that the scalar θ, an overall 'goodness of fit' indicator is calculated as

$$\theta = (1/M) \cdot \sum_r \left(\frac{|\sum x_{l,r} w_l - C_r|}{C_r} \right) \tag{6.5}$$

and set a first threshold value equal θ_{min} to the initial θ value. The *IPU* algorithm then proceeds as follows:

1. For each population characteristic type r, let $\mathbf{V}_r$ be a vector of all household IDs that correspond to components that are non-zero in the $r-th$ column of $\mathbf{X}_{N_h*M}$ such that entry v_{sr}, is an index corresponding to non-zero elements in the $r-th$ column. Update weights w_l such that new weights

$$w^{\star}_{v_{sr}} = \phi w_{v_{sr}} \tag{6.6}$$

where the adjustment factor ϕ is calculated as

$$\phi = C_r / \sum_s (x_{v_{sr},r} * w_{v_{sr}}) \tag{6.7}$$

2. Based on the weights that have been adjusted with respect to all characteristic types, re-calculate the new 'goodness of fit' indicator θ_{new} using Equation 6.5 and the corresponding change in fit $\theta_{diff} = |\theta_{new} - \theta|$ and proceed to assess the improvement in 'goodness of fit'.

3. If $\theta_{new} < \theta_{min}$, update $\theta_{min} = \theta_{new}$ and store corresponding current weights $w_l^\star$ accordingly. Otherwise, continue to the next Step.

4. Iterate until a pre-specified convergence criterion. Based on the improvement in 'goodness of fit' θ_{diff}, if $\theta_{diff} > \varepsilon$, a pre-specified small positive threshold value, repeat steps 1-3. Else, convergence has been reached and the calibrated final weights are obtained.

Once the calibration of weights has been completed, households are then drawn from the household sample data set with respect to the assigned weight from the *IPU* procedure. Readers are referred to Ye et al. (2009) for details of the procedure and also for a proposed extension of the procedure to match household level constraints perfectly.

6.4 Measures of Comparison

The three main procedures of population synthesis are implemented and applied on the Flemish data set available in this research. The algorithms are all coded in the *R* software and efficiency is maximized as much as possible for better comparison of the procedures. For all algorithms, households are drawn here using a Monte Carlo based procedure following the computed probability of selection. Assessing viability of these strategies as a way of generating valid synthetic population data is important. Assessment is thus made for each individual method and then further comparisons are made to determine the *best* approach.

To compare the joint distribution of synthetic data to that of *actual* data, a distance based measure; the Average Absolute Relative Difference (*AARD*) is used as an indication of the 'goodness of fit' for the population synthesis procedures. Consequently, comparison is then made between the procedures. The *AARD* value is calculated using a formula similar to that in Equation 6.5. Given the final synthetic joint distribution obtained at the end of each synthesizing procedure and the known target joint distribution with respect to the control variables, the *AARD* for the household-level is computed as

$$AARD = (1/K_h) \cdot \sum_{i=1}^{K_h} \left| h_i^{new} - h_i \right| / h_i \tag{6.8}$$

where all notations are as defined before. The *AARD* is calculated in a similar way for the person-level joint distribution. Values of *AARD* that are close to zero indicate a *good* fit. Furthermore, to test the null hypothesis that the estimated joint distribution matches the target joint distribution, the well known Chi-square test (Agresti, 2002) is also applied.

For further validation, we examine how well the synthetic population matches other variables which were not controlled for explicitly. In this case, the distributions of some unconstrained variables are explored versus the real data both at the household and the person level.

6.5 Discussion and Conclusion

This Chapter has considered the problem of creating micro-level synthetic data for a population. The framework for generating synthetic data for a target population has thus been presented in the setting where aggregate population data and detailed sample data is available for some variables of interest. Three different procedures have been discussed in detail and formalizations of the methods have been provided. Further details on the implementation of the procedures are given in Appendix A. These population synthesis methods are part of a larger detailed synthetic population generation model designed for Flanders as shown in Figure A.1 in the Appendix. These models have been implemented in the R statistical computing platform.

Conceptually, the methods are also interesting to compare. The *IPU* is capable of reaching a better optimized status than Guo and Bhat (2007)'s procedure. Guo and Bhat (2007)'s algorithm operates by randomly drawing households from a sample into the current synthetic population that is continuously updated based on the household and person-level joint distribution targets. The implication of this approach is that if the distributions of persons in the currently selected household (that satisfies the household-level selection criterion) result in exceeding the targets in any cell at the person-level, the household is not included in the synthetic population. This household is also discarded from the sample before re-computation of selection probabilities and it is not considered again for further selection as it would never pass the selection criteria from this point onwards. This trial-and-error approach is liable to bias in one direction with a consequence of high computation time. This is because in the beginning, the selection criteria are easier to satisfy and thus initially selected households are less likely to be rejected than latter households which may have been more helpful in reaching a globally optimized solution. For large target populations with joint distributions that contain sufficiently large frequencies in each cell, this potential bias is only likely to arise towards the end of the procedure when some cell targets in the synthetic joint distribution have been (or are close to being) reached. Nevertheless, larger households are more likely to be rejected more often towards the end of the procedure as it becomes more difficult for each household member to satisfy the selection criteria. On the other hand, the *IPU* operates as a mathematical programming method that directly solves an optimization problem thus being able to reach a better optimum status than Guo and Bhat (2007)'s algorithm. Matching both household and person-level targets should ideally be formulated as an integer programming problem but this is limited by computational complexity. The *IPU* algorithm tackles this complexity by converting the integer programming problem into a continuous programming problem by using continuous weights instead of a indicator '0-1' variable of whether the household is chosen or not. The continuous programming problem is then solved by a heuristic method and a Monte Carlo procedure is applied to randomly choose households according to probabilities computed from continuous household weights. The *IPU* is also bound to be faster than Guo and Bhat (2007)'s method as it does not alternate computation of selection probabilities with household selection. Probabilities are computed only once. Guo and Bhat (2007)'s algorithm has a merit of preserving the survey design of the base sample used for household selection as it uses the household weights in the computation of selection probabilities. This is opposed to the *IPU* algorithm which initializes the weights to 1 at the start of the procedure. Considering the case of extending the algorithms to more than two analysis levels, it appears easy and natural to include further levels, with the *IPU*. Further levels such as vehicle, and trip-level distributions may be of interest. In such cases however, it seems that Guo and Bhat (2007)'s algorithm, if applicable, will be faced by prohibitively very high computation time. It is also anticipated that if application of the methods is required at much lower disaggregate levels, Guo and Bhat (2007)'s algorithm would require much more computation time as compared to the *IPU* method. Calibration of weights by each small geographic region and then selecting households for each region would presumably cost less time as compared to running the complete trial-and-error method of Guo and Bhat (2007) by each small region.

The application and further discussions on the methods follow in the next Chapter.

7 Creating Synthetic Populations for Flanders

7.1 Introduction

In Chapter 6, we presented and discussed the framework representing different methodologies for generating synthetic populations. In this current Chapter, these methodologies are applied to the setting of Flanders to generate synthetic populations. Consequently, the methods are investigated in a real life setting and performance of the methods is also assessed. To date, the three approaches have never been compared in the literature so far within the same setting of application. Moreover, issues regarding comparison of computational performance are hardly discussed in literature. For most practitioners, population synthesis methods are nothing but a myth that they are quite reluctant to explore. The benefits of the methods are quite un clear and the decision on which method is preferable is not clear-cut. It is the goal of this Chapter to fill this gap by applying the methods on Flemish data with an anticipation of consequently providing guidance, or at least insight to practitioners and researchers on the performance and choice between the methods.

To recapitulate, in creating a synthetic population, the joint distribution of a target population is initially estimated using *IPF*. The estimation is performed separately at the household and the person level based on some control variables of interest with respect to the data available. Based on the estimated joint distributions, different procedures are applied to draw households from a sampling distribution to generate the intended synthetic population. Consequently, different synthetic populations are generated and the results are presented and compared to examine the quality of the generated synthetic data and to offer a means of evaluation of the different population synthesis techniques.

Within the global research focus *FEATHERS*, it is of interest to build scenarios for the population over several future years in the ultimate effort of studying travel behavior. In this thesis, the recreation of the population of Flanders for the year 2001 can be investigated based on the rich data available with respect to this year. For later years however, less and less data are available. Therefore, it is of interest to investigate the creation of a synthetic population in presence of limited data for a later year. The best choice for this application was that of the year 2007. For the year 2007, data are accessible for a sample survey, *FHTS'07*, which are nevertheless assumed not to be available at the modeling stage. The data are rather withheld for validation purposes. Synthesizing the population of 2007 should provide an initial indication on the quality of synthetic data that can be created presently for future years (e.g. 2021). To be more precise, the synthetic population of 2007 would be created based on the data of 2001 assuming no sample data are available for 2007. If the validation of the synthetic population of 2007 with the *FHTS'07* is positive, then there can be confidence that synthetic populations created for future years, for which no concrete validation data are available, can also be provide reliable results.

The first goal is to examine the performance of each algorithm independently, as well as in comparison with the other procedures. To achieve this, generation of the synthetic population of Flanders 2007 is utilized. In conducting the comparisons, focus is directed on control variables as well as variables that are not directly controlled for. The approaches considered here have never been compared in an application setting in the literature so far. After the evaluation process, the second goal is then to use the *best* procedure to generate the required synthetic populations of Flanders for the years 2001, 2007 and 2021. The discussion of the results is thus organized into those pertaining to the creation of synthetic populations through different procedures, using a comparative approach and the results on the synthetic population of Flanders across years, respectively. It is anticipated that one of the methods will merge out superior, outperforming the others. Based on the conceptual comparison conducted in the previous Chapter, it is hypothesized here that the *IPU* will be superior to the other methods considered.

7.2 Data

The data available in this research have been described in Chapter 2. Keeping the data requirements [see Section 6.2] in focus, the data used in the different operations are highlighted here.

To generate the synthetic population of Flanders for the year 2007, the *SEE'01* data [see data in Table B.1 and Table B.2] are used together with the marginal data shown in Table B.3 and Table B.4. In order to select households (and consequently individuals) a sample from the Flemish population for the year 2001, here after referred to as *FS* was drawn from the Census data to provide a basis for the task of selection of households into the synthetic population. The decision to use a sample from the data, instead of the entire *SEE'01* data was motivated by resource restrictions. It is not feasible in most software to model based on the entire population data set. Many statistical software are only capable of accommodating a given maximum size of data. In most applications, micro data are frequently available from samples and data for the entire population are are extremely rarely available. Working with the entire population data set would be highly computer intensive, requiring an enormous amount of time for the completion of the procedures. In R, the software used for implementation of the algorithms, it is currently not possible to handle the entire *SEE'01* data and run all necessary computations on it. Taking a representative sample from the population file therefore seemed the reasonable choice here. The sample design used to select the units of the *FS* closely follows that used in the *FHTS'00*. A stratified sampling design was used in which provinces, household size and the age of the household head were utilized as stratification variables. The procedure then selects independent samples from these strata using a simple random sampling method. The *FS* covers all variables available in the *SEE'01* and comprises of 4005 households with a total of 7970 individuals. The *FS* provides a rich sample which contains data at the household-level as well as data on the person-level for each member of a household. A sample survey data is also available from a travel survey of 2007 (*FHTS'07*). However, besides household-level data, *FHTS'07* data contain person-level data on only one member of each household. The *FHTS'07* survey will provide a basis of comparison with the synthetic data for validation purposes. This is especially for the non-control variables where true data are not available for the population.

In the procedure for generating the synthetic population of Flanders for 2001, the *FS* sample is also used in conjunction with the known joint distributions based on the *SEE'01* data shown in Table B.1 and Table B.2. In an alternative validation procedure, the *FHTS'00* was used instead of the *FS* sample.

For the synthetic population of Flanders for the year 2021, again the the *SEE'01* data as shown in Table B.1 and Table B.2 are used. The data are used together with the margins for the year 2021 obtained from the website of Studiedienst Vlaamse Regering (2009). These margins can also be seen in Table 7.14 and Table 7.15 under the column labeled 'Flemish pop. 2021'. It was of interest to incorporate car ownership as a control variable in generating the synthetic populations. However, data on this attribute were only available for a few years. A model was therefore developed as seen in Chapter 5 to predict car ownership. The resultant data were then subsequently used in the procedures.

7.3 Creating Synthetic Populations through Different Procedures: A Comparative Approach

In this Section we present the results of the synthetic populations of Flanders for 2007 that are generated based on different procedures, within which, validation of the procedures is handled in great detail. In the latter effort, we also discuss some results of the synthetic population of Flanders of 2001 that is generated based on different sample data sets (*FS* and *FHTS'00*). Further results on the stability of results from the procedures are also finally presented.

7.3.1 Estimation of the joint distribution

To estimate the target joint distributions for Flanders in 2007, the *SEE'01* joint distributions were updated using *IPF* based on the population marginal values of Flanders for 2007. This was conducted both at the household and the person-level. Results were obtained to give an insight on the computational performance when updating the joint distribution with respect to the control variables using the *IPF* procedure. Here, 2.5 million households comprising of the household heads of 18 years or older are synthesized for the Flemish population of 2007. At the individual-level, a total of 6.1 million is synthesized. The performance tests were run using the R software on an Intel Pentium M computer with a 1.73GHz processor and 0.99GB of memory. The stop criterion was set to a very small value of 10^{-6}. The implemented *IPF* algorithm is very fast and runs in only a few seconds (1-3s) with convergence achieved in only 16 and 6 iterations at the household and individual-level respectively. These computation efficiency results are line with literature. A good example is that of Arentze et al. (2007) who developed a method based on *IPF* for generating synthetic populations. In an experiment in Arentze et al. (2007) conducted to illustrate some properties of *IPF*, it was reported that convergence was reached after 17 iterations. Moreover, Beckman et al. (1996) also notes that, in practice, the *IPF* procedure converges in 10-20 iterations.

The final results of the *IPF* routine lead to two estimated joint distributions for the Flemish population of 2007 at the household and individual-level. These joint distributions are then used as input for the next stage of household selection in synthesizing the population. For illustration purposes, let us consider the task of estimation of the household level joint distribution of 2007. To achieve this task, the household level marginal values in Table B.3 are used to update the joint distribution in Table B.1 using the *IPF* procedure. Table 7.1 shows the results of the *IPF* routine in estimation of the household joint distribution of Flanders in 2007. Results are shown for the first, second and third iteration of the routine, respectively. The first three columns show the different levels of the control variables that form a given cell/group of the joint distribution. For instance, the first cell comprises of single households with no car and whose heads are aged between 18-59 years. Column four displays the starting cell counts before any updating is conducted. At the first iteration, adjustments are made by each control variable. Here, for each variable in turn, adjustment factors [see Equation 6.3] are calculated at each level of the variable and the corresponding cell values are adjusted accordingly. At the end of the first iteration, all the adjustments have been made for all the three variables. However, convergence has not yet been achieved and therefore further updating is necessary. At the second iteration, the procedure is repeated, updating all cells considering one variable at a time. The results at the end of the iteration are shown under the corresponding column. Analogously, further results are obtained for iteration three. Iteration continues until convergence.

The final results obtained after convergence are displayed in the last column of Table 7.2. Here, the findings now reveal the entire joint distribution of Flanders for the year 2007. It is observed that for instance, group 1 (single households, no car and household heads aged 18-59) consists of 101,602 households. For the group of households of size 4, without car and with household heads aged 60 years and above (group 8), 1,484 households are estimated to be in the population of 2007 as opposed to the earlier value of 2,125 in the year 2001. From Table 7.2, when the cell values in the final results of the *IPF* procedure are summed up to obtain the marginal distributions by each control variable, it is observed that they match, almost perfectly, those shown in Table B.3. This is not surprising as it is the goal of the *IPF* procedure. Also, the total summation of the target population is preserved at each iteration. The *IPF* procedure is also applied at the person-level to obtain similar corresponding results. Here, the size of the population estimated through the *IPF* procedure is large, with large cell sizes as well. This is also also probably contributes also to the high level of accuracy of the estimates. Wong (1992) investigated the reliability of *IPF* when an original matrix of control variables is derived from a sample. Using sample populations from Public-Use Micro data Samples(PUMS) and statistical test of total absolute error, the random error effect was found to be influenced by the size of both the sample and matrix. Wong (1992) also established that increasing the sample size increased the accuracy of estimates derived using *IPF*. Moreover, this effect was more pronounced for larger matrices.

7.3.2 Application of the Procedures

The variables used in the step of estimation of the joint distribution are now used here as control variables. Thus, at the household-level, the control variables include: *HHDER-AGE* which has groups 18-59, and 60plus; *HH-AUTO* a binary variable; *HHSIZE* for which households of size 4 and above are grouped together leading to 4 categories for this variable. This aggregation is done to avoid zero-cell values for corresponding groups in the *FS* that is used as the basis for household selection. Moreover, re-grouping *HHSIZE* here is a reasonable choice as the population itself contains very few households larger than a size of 4. In general the population consists of very few households that are larger than 5 household members. At the individual-level, *P-GENDER* and *P-AGEGRP* are controlled for with their categories as defined before. There are thus 16 and 38 cells at the household level and the person-level joint distributions respectively.

Overall performance

Three procedures: *P1* following the conventional approach, *P2* following Guo and Bhat (2007)'s procedure and *P2* the *IPU* procedure by Ye et al. (2009) are applied to generate synthetic populations for Flanders for the year 2007. The performance results for the procedures are shown in Table 7.3. The performance tests were run on an Intel Xeon with a 3.00GHz processor and 16.0GB of memory. Overall, all the three synthetic population generation algorithms are quite efficient with respect to the reasonable computation time used. These results show that *IPU* is clearly more efficient as compared to the other approaches. It takes just over a minute to obtain the entire synthetic population of Flanders using *IPU* as compared to the approximately 6 hours required by Guo and Bhat (2007)'s algorithm. The conventional procedure on the other hand also requires just a few (less than 5) minutes to execute this job. The order of performance for the procedures remains the same when only proportions of the target population size are generated. It is not surprising though that the conventional procedure is much more efficient in speed than Guo and Bhat (2007)'s algorithm since the conventional algorithm does not consider person-level information so the computation time that would have been spent on that task is saved. In the author's opinion, the difference between Guo and Bhat (2007)'s approach and the *IPU* is a relatively large one in terms of computation time. This therefore highly favors the *IPU* method since in practice, when one needs to apply a model, computation efficiency is very important. The computation efficiency difference had been however foreseen in the conceptual comparison of the methods in Chapter 6, where the trial and error approach of Guo and Bhat (2007) had been anticipated to result in loss of time. Comparing computational efficiency across studies conducted by different researchers is not always straight forward as they are normally performed under different settings. Whereas Guo and Bhat (2007) does not give a clear indication of the computational efficiency of their method, Ye et al. (2009) note that in the application of the *IPU* method, the total processing time in a single-core configuration was approximately 16 hours, in the setting which involved creating synthetic populations for the entire Maricopa County region of Arizona for all 2088 blockgroups. Ye et al. (2009) used a Dell Precision Workstation with Quad Core Intel Xeon processor was used to run the entire algorithm. The algorithm was coded in Python a dynamic open-source object-oriented programming language and the data was stored and accessed using MySQL a commonly used open source database solution. The code was parallelized to take advantage of the multiple cores in the processor. If data would be available to permit implementation of the algorithms considered here by smaller regions such as municipalities, it is clear that the procedures would require more computation time than reported in this thesis. However, it is reasonable to expect that the order of performance of the procedures would remain the same. Moreover, the *IPU* would be expected to achieve the target number of units at a disaggregate level in a reasonable amount of time.

The *IPU* is also able to achieve very closely the target total number of households and the total number of persons. On the contrary, it appears to be more difficult to achieve the target total number of households and persons with Guo and Bhat (2007)'s procedure. This had also been foreseen in our conceptual analysis where it was noted that achieving an optimal solution is not straightforward with this method. This disadvantage has been more-or-less reflected in *P2-2* and *P2-3* columns of Table 7.4. When household *AARD* is fixed at 0, Guo and Bhat (2007)'s algorithm can gain person *AARD* at 0.082 and 0.099, which is only slightly better than 0.111 under the situation that person-level distribution is not controlled. The approach of Guo and Bhat (2007) involves discarding of households making it virtually infeasible to achieve the target number of households and persons. Achieving an optimal solution is not straightforward with Guo and Bhat (2007)'s algorithm.

Table 7.4 shows the 'goodness of fit' results obtained by application of the three household selection algorithms to generate synthetic data. The results for the conventional procedure (*P1*) are shown in the second row. Guo and Bhat (2007)'s algorithm is applied for different experiments allowing for different levels (0%, 5% and 10%) of *PDS* from the target household and person distribution sizes. The corresponding results are shown by row, from *P2-1* to *P2-9* for nine experiments. The last row displays the *IPU* algorithm results. No percentage deviation needs to be allowed for the latter procedure since the method is able to perfectly match household targets by design. The results for calibration of household weights to match both household and person-level joint distributions using *IPU* were achieved in 290 iterations. Overall the fit for all the procedures is good since the *AARD* values are considerably close to zero. The *IPU* was implemented to match the household-level distribution perfectly hence the observed '0' *AARD* value at this level is expected. Even then, at the person-level, the *IPU* still provides the best fit of 0.01. On the other hand, *P1* which also matches the household-level distribution perfectly gives the worst fit of 0.11 at the person-level; a relative bias of the individual level distribution. Application of another distance-based measure - the Standardized Root Mean Square Error (*SRMSE*), yields similar findings in assessing the 'goodness of fit' of the selection procedures.

For the second procedure (*P2*), using different implementations (P2-1 to P2-9), the *AARD* value is generally below 0.1 for all of these procedures. Allowing for no (0% *PDS*) oversampling at both household and individual-level yields relatively moderate *AARD* values at both levels while yielding a fairly accurate average household size of the population that is close to the true value of 2.40. However, when the PDS_{PER} is allowed to increase slightly, say up to 5%, the target household-level joint distribution may be achieved perfectly at a small cost of a slight more distortion at the person level. It appears that allowing for very low (or no) PDS_{PER} while allowing for some PDS_{HH} (i.e. scenarios *P2-4*, *P2-5* and *P2-7*) results in moderate *AARD* values at both levels. However, the average household size further becomes lower than the expected in these cases. As thus, the decision of how much deviation should be allowed at what level thus remains up to individual researchers to make based on their research goals. An argument can be put forward that in travel demand analysis in general, it is individuals who actually travel and not households. If higher importance is attached to satisfying the target individual-level distribution as compared to the household-level distribution, it appears suitable to permit for higher PDS_{HH}, say, 10% with no deviation allowance at the individual-level. In this case, as more households are selected, the individual-level distribution which is not allowed to deviate from the target levels gets more refined. If interest is the opposite, then some PDS_{PER} could be allowed at the individual-level with no deviation allowance at the household-level. Nevertheless, irrespective of the alternative permitted *PDS* values, the target distributions are fitted relatively well with the overall $AARD$ between the synthetic and the true data quite close to 0. All procedures are able to yield values that are relatively close to the true average household size of 2.40 as demonstrated in Table 7.4. This is the true value based on the population register of Flanders for the year 2007. However, the *IPU* is more precise in this estimate. Guo and Bhat (2007)'s study also investigated the effect allowing for different $AARD$. In their case study, it was noted that a higher $AARD$ (10%) appeared to strike a better balance at satisfying both the household- and individual-level multi-way distributions than allowing lower values of $AARD$ (0% and 5%). However, the choice between allowing for 0% or 5% $AARD$ was not definite. The study did not investigate the possibility of allowing different $AARD$ values at the household and person-level in a given experiment but it was noted that additional validation analysis was needed to better understand the sensitivity of the algorithms performance on $AARD$ values. Therefore, more has been learned from the current study that separately allowed for different $AARD$ values at the household and person-level with an investigation on the effect on the average household size of the generated synthetic populations.

Table 7.1: Results of the cell counts for the *IPF* iteration routines in estimation of the household-level joint distribution of Flanders 2007

Control variables			Flanders 2001	*IPF* iteration			
HH-AUTO	*HHSIZE*	*HHDER-AGE*	Known counts	First	Second	Third	...
no	1	18-59	104784	106476	106476	103057	...
no	1	60plus	208935	199342	199342	192961	...
no	2	18-59	35966	32877	32877	31365	...
no	2	60plus	78607	67467	67467	64371	...
no	3	18-59	19177	15897	15897	14799	...
no	3	60plus	9786	7617	7617	7091	...
no	4	18-59	10849	8905	8905	8207	...
no	4	60plus	2125	1638	1638	1510	...
no	5	18-59	4280	3511	3511	3241	...
no	5	60plus	820	632	632	583	...
no	6	18-59	1539	1289	1289	1195	...
no	6	60plus	404	318	318	295	...
no	7	18-59	621	542	542	506	...
no	7	60plus	181	148	148	139	...
no	8	18-59	259	218	218	204	...
no	8	60plus	98	78	78	73	...
no	9	18-59	121	95	95	89	...
no	9	60plus	45	33	33	31	...
no	10+	18-59	196	75	75	72	...
no	10+	60plus	393	141	141	136	...
yes	1	18-59	243673	305219	305219	315167	...
yes	1	60plus	117063	137675	137675	142177	...
yes	2	18-59	339651	382719	382719	389529	...
yes	2	60plus	354584	375143	375143	381857	...
yes	3	18-59	310709	317489	317489	315325	...
yes	3	60plus	73628	70640	70640	70165	...
yes	4	18-59	323720	327532	327532	322056	...
yes	4	60plus	21575	20496	20496	20155	...
yes	5	18-59	107521	108731	108731	107068	...
yes	5	60plus	9571	9088	9088	8950	...
yes	6	18-59	26424	27273	27273	26989	...
yes	6	60plus	4571	4430	4430	4384	...
yes	7	18-59	6253	67313	6731	6702	...
yes	7	60plus	1791	1810	1810	1802	...
yes	8	18-59	2280	2368	2368	2364	...
yes	8	60plus	746	728	7278	726	...
yes	9	18-59	965	937	937	936	...
yes	9	60plus	332	303	303	302	...
yes	10+	18-59	979	462	462	473	...
yes	10+	60plus	1392	616	616	632	...
Total			2426614	2547686	2547686	2547686	...

Table 7.2: Results of the cell counts for the final *IPF* iteration routine in estimation of the household-level joint distribution of Flanders 2007

Control variables			Flanders 2001	Flanders 2007
HH-AUTO	*HHSIZE*	*HHDER-AGE*	Known counts	Estimated counts
no	1	18-59	104784	101602
no	1	60plus	208935	190877
no	2	18-59	35966	30734
no	2	60plus	78607	63288
no	3	18-59	19177	14501
no	3	60plus	9786	6972
no	4	18-59	10849	8041
no	4	60plus	2125	1484
no	5	18-59	4280	3175
no	5	60plus	820	573
no	6	18-59	1539	1171
no	6	60plus	404	290
no	7	18-59	621	496
no	7	60plus	181	136
no	8	18-59	259	200
no	8	60plus	98	71
no	9	18-59	121	88
no	9	60plus	45	31
no	10+	18-59	196	71
no	10+	60plus	393	133
yes	1	18-59	243673	317269
yes	1	60plus	117063	143607
yes	2	18-59	339651	389738
yes	2	60plus	354584	383348
yes	3	18-59	310709	315480
yes	3	60plus	73628	70436
yes	4	18-59	323720	322182
yes	4	60plus	21575	20231
yes	5	18-59	107521	107113
yes	5	60plus	9571	8983
yes	6	18-59	26424	27002
yes	6	60plus	4571	4401
yes	7	18-59	6253	6707
yes	7	60plus	1791	1810
yes	8	18-59	2280	2366
yes	8	60plus	746	729
yes	9	18-59	965	937
yes	9	60plus	332	304
yes	10+	18-59	979	474
yes	10+	60plus	1392	635
Total			2426614	2547686

Table 7.3: Computation performance results for the different synthesis algorithms

Population	Procedures	No. of Persons	No. of Households	Duration
5% of Flanders				
Synthetic	*P1*	302484	127386	<1min
	P2	297433	125498	17 mins
	IPU	305438	127384	0.5mins
Entire Flanders				
Synthetic	*P1*	6049480	2 547686	<5min
	P2	5949633	2509970	5.7 hrs
	IPU	6117783	2547686	1.5mins
True		6117440	2547686	

Table 7.4: Comparison of the performance of different population synthesis Procedures

Procedure	Controlled distributions		Allowed *PDS* values(%)		*AARD*	*AARD*	Average
	Household	Individual	PDS_{HH}	PDS_{PER}	(household- level)	(Person-level)	household size
P1	Yes	No	0	N/A	0	0.111	2.37
P2-1	Yes	Yes	0	0	0.016	0.050	2.37
P2-2	Yes	Yes	0	5	0.0	0.082	2.37
P2-3	Yes	Yes	0	10	0.0	0.099	2.37
P2-4	Yes	Yes	5	0	0.029	0.039	2.36
P2-5	Yes	Yes	5	5	0.029	0.038	2.36
P2-6	Yes	Yes	5	10	0.050	0.088	2.37
P2-7	Yes	Yes	10	0	0.035	0.022	2.36
P2-8	Yes	Yes	10	5	0.047	0.064	2.35
P2-9	Yes	Yes	10	10	0.084	0.095	2.37
IPU	Yes	Yes	0	0	0	0.010	2.40

Procedure P1 corresponds to an implementation of the approach proposed by Beckman et al., 1996 (Beckman et al., 1996), procedures P2-1 to P2-9 correspond to implementations of Guo and Bhat 2007's (Guo and Bhat, 2007) algorithm (with different levels of percentage deviations from target size allowed for) and the IPU is the procedure proposed by Ye et al., 2009 (Ye et al., 2009).

The performance of the procedures can be further viewed graphically in Figure 7.1, which shows plots of the cell proportions for the synthetic versus the target joint distributions by the different selection procedures. The distributions are shown both at the household, in the left panel and person-level, in the right panel. The points represent the synthetic cell proportions generated by the respective methods. For a perfect fit, all the points would lie on the line. By design, the *IPU* procedure and *P1* are implemented here to match the household-level distribution perfectly, as previously noted, thus it is not unexpected to see a perfect fit at the household-level as observed in the figure. More still, *IPU* clearly provides the best fit at the person-level as well. In contrast, considerable distortion is observed for *P1* at the person-level. For purposes of the subsequent investigations, the result of procedure *P2-1* is considered sufficient regarding Guo and Bhat (2007)'s method. Therefore, for clarity purposes, from now on when we refer to Guo and Bhat (2007)'s method as *P2*, we refer to the setting of procedure *P2-1*.

Marginal distribution of control variables

The accuracy of the three procedures in estimating the distributions of some variables of interest was also investigated. Firstly, the results of the control variables are examined, followed by some interesting variables that were not explicitly controlled for due to lack of information at the marginal population level. Table 7.5 shows the marginal distributions of the control variables at the household and person level. The distributions with respect to the sample survey (*FHTS'07*) conducted in 2007 in Flanders are reflected in the second column whereas the true distributions from the population register of Flanders for 2007 are shown in the third column. One can observe that, the actual distributions are quite well preserved across all procedures for all variables. As already indicated, the *IPU* and *P1* procedures show very accurate representation of the household distributions when the resultant synthetic data are compared to the population data. Considering the distributions from the independent sample (*FHTS*), it is remarkable that similar results can be easily obtained from a synthetic procedure which requires much less resources. Given that conducting sample surveys is costly, time consuming and is faced by other problems, the success of population synthesizing procedures is paramount. Government bodies and other institutions can potentially save

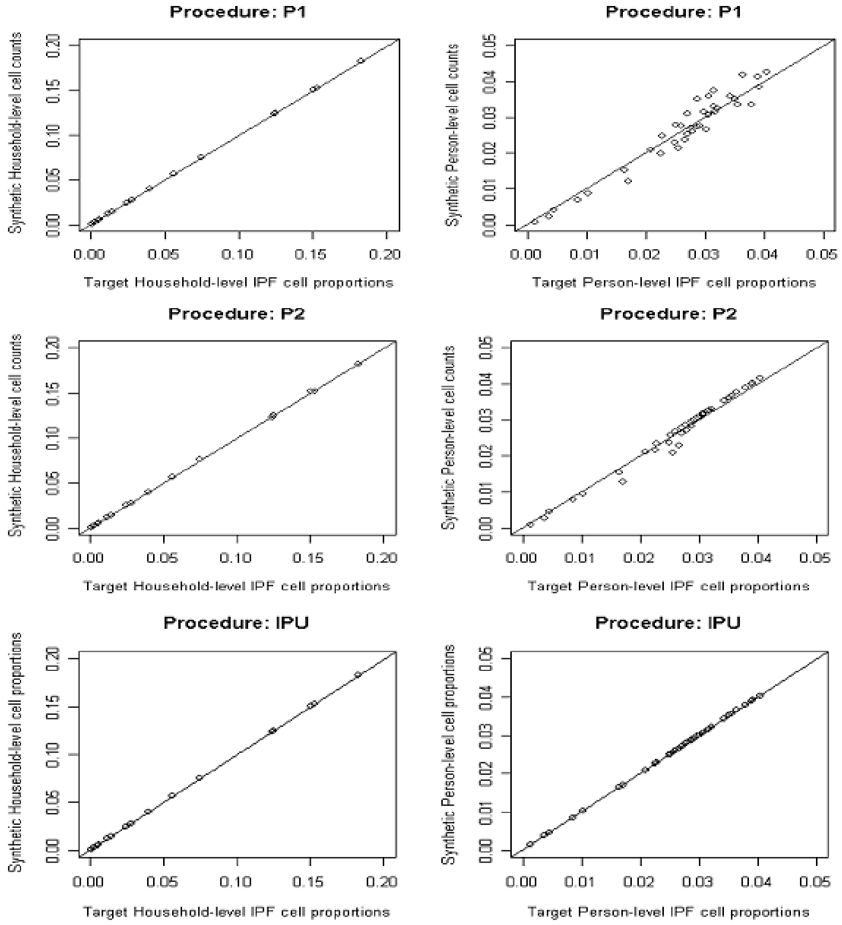


Figure 7.1: Plot of the Synthetic versus the Target Joint Distributions by Different Household Selection Procedures for Population Synthesis shown at both the Household and Person-level. Procedure P1 corresponds to an implementation of the approach proposed by Beckman et al. (1996), procedure P2 correspond to Guo and Bhat (2007)'s algorithm and the IPU is with respect to the procedure proposed by Ye et al. (2009). Plotted points represent cell proportions of the synthetic joint distribution for control variables.

vast amounts of resources through use of data synthesizing approaches. So far, it can be argued that if all the variables in the required synthetic data sets can be controlled for, the *IPU* method appears to be a definite preferable choice compared to the other methods. It is anticipated that this result would hold even in the the case where the population is drawn for smaller geographic regions. For each small geographic region, the distributions would be controlled with respect to the control variables and thus the distributions would accordingly be refined at this level.

Table 7.5: The distributions of the control variables

Variables	Actual data			Synthetic data 2007		
	FS	*FHTS'07*	Flemish pop. 2007	*P1*	*P2*	*IPU*
Person level						
Gender						
Male	49.38	49.02	49.31	49.59	49.61	49.40
Female	50.62	50.98	50.69	50.41	50.39	50.60
Age categories						
34less	41.73	36.48	41.35	41.46	40.47	40.39
35-54	29.94	31.79	29.82	30.24	30.67	29.93
55-75	21.90	23.61	21.80	22.18	22.18	22.21
76plus	6.43	8.12	7.02	6.13	6.67	7.47
Household level						
Age group -Oldest householder						
18to59	63.21	63.88	64.74	64.74	64.40	64.74
60plus	36.79	36.12	35.26	35.26	35.60	35.26
Household size						
1	27.75	28.00	29.57	29.57	29.70	29.57
2	33.34	33.55	34.04	34.03	34.12	34.03
3	17.05	16.08	15.99	15.99	15.94	15.99
4+	21.86	22.36	20.40	20.40	20.23	20.40
Possession of at least 1 car						
Yes	80.63	81.79	83.36	83.36	83.29	83.36
No	19.37	18.21	16.64	16.64	16.71	16.64

Procedure P1 corresponds to an implementation of the approach proposed by Beckman *et al.*, 1996 (Beckman et al., 1996), procedure P2 correspond to Guo and Bhat 2007 (Guo and Bhat, 2007)'s algorithm and the IPU is with respect to the procedure proposed by Ye *et al.*, 2009 (Ye et al., 2009)

Marginal distributions of non-control variables

In order to get deeper insight into the performance of the methods under review, further investigation is conducted. The population synthesis procedures also allow generation of more data besides those based on the household and person-level control variables. These are characteristics of households and persons that come with the households selected from the base sample into the synthetic population. Although we do not advocate for the such 'free' data to be 'inherited' literary and used for various purposes as any other actual set of data, we are of the view that if non-control variables can be shown to be replicative of actual data, these data could be useful.

To examine how well the synthetic population matches other variables which were not controlled for explicitly, the distributions of some unconstrained variables were explored versus the real data both at the household and the person level. This is also one method of validating the schemes for creating the synthetic populations. The distributions that provide the real data are based on *SEE'01*, *FS* and *FHTS* shown in the first columns. All these data sets aim at representing Flanders since they are taken for the same population. Moreover, the *SEE'01* should represent the true data of 2001 since it arises from a population census. The *FS* is used as the basis for household selection in the population synthesis procedures and *FHTS* is the survey sample data set intended for explicit validation. The distributions of the synthetic populations of 2007 are not expected to vary significantly from those of the *SEE'01* except if time plays a very important role here. Nevertheless, assuming that the *FHTS* sample is a representative one, since the data are synthesized for the Flemish population for the year 2007, this survey data set should provide the most ideal basis of comparison. No population data are available for comparison for the year 2007.

Table 7.6 shows the distributions of some household-level variables that were not explicitly controlled for during the process of population synthesis. The results for procedures *P2-6* to *P2-9* are not shown as they are not very different from those observed in procedures *P2-1* to *P2-5*. While possession of a car was controlled for, the total number of cars in a household was not. Impressively, most of the distributions for the synthetic data are generally reasonably close to the 'true' values from real data based on the variables under consideration for all procedures. No particular distortions are observed in the data. A consideration of the distributions from the *FHTS* for most variables reveals that synthetic data can be accurate in general with respect to the procedures considered here. However, the number of households without bicycles and those with one bicycle tends to be overestimated in the synthetic data as compared to the values provided by the *FHTS* sample data. Of course, it has to be kept in mind that the *FHTS* is a sample, implying that it is only one of the many possible random samples that could be extracted from the population. It is therefore possible for the sample distribution to deviate from the 'real' distribution. Nevertheless, in practice, what is frequently available to researchers is a sample rather than the data on the entire population.

Table 7.6: The distributions of some household-level non-control variables

Household-level	Real data			Synthetic data 2007						
variables	*SEE'01*	*FS*	*FHTS'07*	*P1*	*P2-1*	*P2-2*	*P2-3*	*P2-4*	*P2-5*	*IPU*
No. of cars										
0	19.75	19.37	18.21	16.64	16.71	16.64	16.64	16.87	16.88	16.64
1	55.36	56.53	53.64	59.23	59.77	59.51	59.13	59.70	59.73	60.16
2	22.15	21.69	24.75	21.77	21.22	21.50	21.84	21.20	21.19	21.15
3+	2.74	2.42	3.39	2.36	2.30	2.35	2.40	2.23	2.20	2.05
No. of bicycles										
0	21.88	22.80	18.56	22.71	22.97	22.57	22.65	23.07	23.22	22.99
1	24.39	24.23	19.38	24.94	24.74	24.74	24.96	24.81	25.01	25.07
2	26.81	26.02	25.22	26.48	26.62	26.56	26.36	26.79	26.44	26.42
3+	26.92	26.96	36.84	25.87	25.66	26.13	26.04	25.33	25.33	25.51
Province										
Antwerp	28.43	28.40	26.06	28.55	28.58	28.52	28.36	28.54	28.57	28.86
Vlaams Brabant	16.98	17.00	17.07	17.03	17.16	17.07	17.35	17.30	17.36	17.08
West Flanders	19.04	19.04	19.09	19.10	18.88	18.86	18.94	18.85	18.84	18.72
East Flanders	23.09	23.06	23.21	22.79	22.95	22.97	22.89	22.95	22.90	22.94
Limburg	12.46	12.50	14.59	12.53	12.42	12.57	12.46	12.36	12.33	12.39

Procedure *P1* corresponds to an implementation of the approach proposed by Beckman *et al.*, 1996 (Beckman et al., 1996), *P2-1* to *P2-5* procedures correspond to implementations of Guo and Bhat 2007 (Guo and Bhat, 2007)'s algorithm (with different levels of percentage deviations from target size allowed for) and the *IPU* is with respect to the procedure proposed by Ye *et al.*, 2009 (Ye et al., 2009).

Table 7.7 shows the distributions of some person-level variables that were not controlled for during the process of population synthesis. Overall, the distributions based on the synthetic data are quite close to the actual distributions for all procedures. Even procedure *P1* which does not involve controlling for person level variables yields reasonable distributions at the individual-level. The distributions for some travel-related continuous variables are also well estimated (across all procedures) with respect to both the mean and standard deviation. The mean distance for work/school trips for persons who either work or go to school is 16km as per the *FHTS'07* data. This mean is relatively accurately estimated as 15km based on the synthetic data from each of the procedures. In addition, using all procedures, the true mean (according to *SEE'01* and *FS*) work hours for the working population is preserved. The distribution of the personal income variable from the *FHTS'07* could not be provided for comparison since the data were collected based on a different set of incompatible categorization. Also data on work hours are not available in the *FHTS'07*. It can be said that in general, the synthetic population represents well the true population based on most of unconstrained household and person-level variables. This is also the case for some other variables (such as the occupational status of persons) that are not shown here. Procedures (*P2-1* to *P2-5*) appear to replicate the original sample data (*FS*) quite well thereby supporting the underlying methodology as a good choice if the sample data that are used as the basis for selecting households for the synthetic population are truly representative of the target population. However, it is also noteworthy that the *FS* that is used as the basis for household selection, much as it is considered here as representative of actual data; it also inhibits some small deviations of distributions from the corresponding population data. It would be expected that these small biases would be transferred to the synthetic data thus attributing some of the inconsistencies observed in the synthetic data to this reason. This underscores the importance of the quality of sample data set that is used in the population synthesis procedure. In general however, it is remarkable that even with these underlying discrepancies the procedures perform as well as they do.

Table 7.7: The distributions of some person-level non-control (unconstrained) variables

Person-level	Real data			Synthetic data 2007						
variables	*SEE'01*	*FS*	*FHTS'07*	*P1*	*P2-1*	*P2-2*	*P2-3*	*P2-4*	*P2-5*	*IPU*
Driver‘s license										
Yes	64.05	63.84	74.23	64.21	64.47	64.50	64.62	64.78	64.86	64.47
No	35.95	36.16	25.77	35.79	35.53	35.50	35.38	35.22	35.14	35.53
Personal income (euros)										
0	24.88	24.99	-	24.35	24.36	24.23	24.18	24.08	24.15	24.84
1-1250	21.71	22.54	-	22.83	22.68	22.83	22.85	22.46	22.49	22.05
1251-2250	43.12	42.59	-	42.84	42.75	42.72	42.91	43.22	43.19	43.03
>2250	10.29	9.88	-	9.98	10.21	10.22	10.06	10.24	10.17	10.08
Mode work/sch. trips										
Slow mode	22.48	24.16	25.56	23.70	23.68	23.84	23.55	23.90	23.95	23.74
Car	61.05	59.88	56.36	60.36	60.09	60.27	60.61	60.14	60.12	60.05
Public	13.26	12.93	16.47	12.91	13.01	12.81	12.83	12.85	12.93	13.18
Other	3.21	3.02	1.6	3.03	3.22	3.08	3.01	3.11	3.00	3.04
Work/school										
Yes	65.87	66.16	66.90	66.32	65.84	66.13	66.18	65.34	65.28	65.02
No	34.13	33.84	33.10	33.68	34.16	33.87	33.82	34.66	34.72	34.98
Mean(standard dev.)										
Dist. work/sch. trips	15.18(19.60)	14.50(18.48)	15.95(20.76)	14.64(18.58)	14.63(18.47)	14.61(18.45)	14.71(18.61)	14.69(18.49)	14.60(18.46)	14.62(18.47)
Work hrs per week	39.07(13.19)	38.96(13.21)	-	39.09(13.26)	38.99(13.23)	39.06(13.22)	39.10(13.30)	38.99(13.25)	39.01(13.24)	38.93(13.16)

Procedure *P1* corresponds to an implementation of the approach proposed by Beckman *et al.*, 1996 (Beckman et al., 1996), *P2-1* to *P2-5* procedures correspond to implementations of Guo and Bhat 2007 (Guo and Bhat, 2007)'s algorithm (with different levels of percentage deviations from target size allowed for) and the *IPU* is with respect to the procedure proposed by Ye *et al.*, 2009 (Ye et al., 2009).

Further validation

In a further effort to validate the procedures, we examined the results at more disaggregate levels. For instance, different variables were explored by province. At this level, the results revealed that the procedures still substantially replicate the actual distributions quite well, with the associations between variables preserved.

Considering the results presented so far, there still remains an open question. Are the results of the synthetic procedure dependent on the particularity of the base sample? Experiments were set up to generate synthetic populations under different conditions to investigate if the conclusions arrived at so far are maintained. In this case, a scenario of generating synthetic populations using another sample data set besides the *FS* was considered. Thus the synthetic populations of Flanders, 2001 are generated based on *the FHTS'00* as well as based on the *FS*. For this experiment, attention is restricted on use of the *IPU* algorithm. In this setting, the *SEE'01* is then used for validation purposes. Before the analyses are conducted, some issues require mention due to the nature of data used. The *FHTS'00* survey covered persons in Flanders aged 6 years and above and data were collected on only 4 members of a household. It would be expected that the synthetic population that will be generated based on *the FHTS'00* would not be exhaustive for all members of the household. Given this, some slightly more observed variation between the *FHTS'00*-based synthetic population and the true data should not be surprising.

Table 7.8 displays the results of the synthetic populations of 2001 for Flanders based on *FHTS'00* and *FS*. The distributions are shown for some control variables and it is clear that no distribution distortion is observed as it was also the case for the synthetic population of Flanders for 2007 based on *FS*. For some non-control variables that are shown in Table 7.9, it is observed that the distributions based on the *FHTS'00* are pretty similar to those in the validation data set *SEE'01*. The only striking divergencies from the true distributions are seen for the variable 'number of bicycles'. However, this deviation had also been existent in the base sample data set *FHTS'00*. In general, some deviation is more or less reflected in the results of the person-level for the *FHTS'00*-based synthetic population as compared to the *FS*-based synthetic population.

Table 7.8: Comparing synthetic data based on different sample data sets - distributions of some control variables

Variables	Real data	Synthetic data 2001 based on	
	SEE'01	*FHTS'00*	*FS*
Person level			
Gender			
Male	49.31	49.17	49.31
Female	50.69	50.83	50.69
Age categories			
34less	41.36	37.93	41.36
35-54	29.82	31.57	29.81
55-75	20.99	23.38	21.84
76plus	7.84	7.12	6.99
Household level			
Household size			
1	27.79	27.65	27.79
2	33.33	33.37	33.33
3	17.03	17.09	17.03
4+	21.84	21.89	21.84
Possession of at least 1 car			
Yes	80.25	80.25	80.25
No	19.75	19.75	19.75

Further experiments on controlling for varying number of variables at the household and/or person-level have also been investigated. The question here of interest was: Does exclusion of some control variable(s) impact the results of the synthesis? If possible it would be interesting to determine which variables are sufficient for a good fit. We considered several scenarios involving excluding one or more control variables at the household and/or the person-level in generating a synthetic population. The general findings were quite intuitive. The results suggested that exclusion of a given control variable tends, in most cases, to slightly distort its resultant distribution in the

Table 7.9: Comparing synthetic data based on different sample data sets - distributions of some non-control variables

Variables	Real data	Synthetic data 2001	
	SEE'01	Based on *FHTS'00*	Based on *FS*
<u>Person level</u>			
Drivers license			
Yes	64.05	67.42	64.10
No	35.95	32.58	35.90
Work/School (Y/N)			
Yes	65.87	64.96	65.67
No	34.13	35.04	34.33
<u>Household level</u>			
No. of cars			
0	19.75	19.75	19.75
1	55.36	52.11	56.17
2	22.15	24.21	21.66
3+	2.74	3.94	2.42
No. of bicycles			
0	21.88	7.75	22.77
1	24.39	23.11	24.08
2	26.81	29.46	26.13
3+	26.92	39.68	27.03
Province			
Antwerp	28.43	30.36	28.48
Flemish Brabant	16.98	16.90	16.83
West Flanders	19.04	19.79	19.01
East Flanders	23.09	20.12	23.24
Limburg	12.46	12.83	12.45

synthetic data population. It therefore appears that using all available data is worthwhile and is crucial in providing a more refined synthetic population. This conclusion is also intuitive.

This study has provided compelling results confirming the positive results of individual research work in literature (Beckman et al., 1996; Guo and Bhat, 2007; Ye et al., 2009; Mohammadian et al., 2010). Although the study achieves the intended goals, there is a limitation that the population is not drawn for small geographic regions. If there is interest in achieving representativeness up to much smaller geographic units, it would be imperative that distributions are controlled at the respective geographic levels of interest. However, this would only be possible if the relevant data on control variables are also available at this level. Besides the necessary aggregate data as indicated in Section 6.2 and 7.2, the sample to be used as a basis for drawing households or units for that matter, should also be detailed to the same level. The implemented methods for population synthesis would then need to be run for each small geographic region. For instance, considering 600 towns in Flanders, 600 runs would then be required to be implemented. It should be feasible to implement such an approach although it is anticipated that the procedures would require much more computation time to complete the population synthesis process. Moreover, issues such as the problem of zero cells would arise and would thus need to be addressed.

Stability of results from the procedures

A crucial point of concern is also the stability of results from different synthetic populations generated using a given method. The Monte Carlo based method used to draw households involves an underlying probabilistic measure and thus this concern naturally arises. To investigate the stability of results from simulation to simulation, 100 experiments were set up to generate different synthetic data sets for Flanders resulting from using each algorithm. The previously discussed results shown in Table 7.6 and Table 7.7 corresponding to Guo and Bhat (2007)'s procedure *P2-1* to *P2-5* also offer an indication of the good stability of results for this approach even when different levels of *PDS* are allowed for.

For the current multiple experiments, results of the resultant joint distributions of control variables are shown for 10 experiments in Table 7.10 and Table 7.11 for the household-level and person-level, respectively based on Guo and Bhat (2007). The runs are performed for 5% of the population. It can be observed that the distributions remain consistent over different synthetic runs. It even appears that some groups are always exactly similar across simulations. For instance at the person-level, groups that relate to age-groups of 5 to 10, These groups are coded with respect to the defined groups in Chapter 2. Table 7.12 further lays out the results for 10 synthetic runs, showing the corresponding computed $AARD$ values. Performing procedure *P2* a hundred times results into a mean $AARD$ value of 0.016 (std. dev.= 0.0003) and 0.052 (std. dev.=0.0005) over all the 100 synthetic populations at the household and person-level respectively. This can also be deduced from the shown results. The standard deviation is very small indicating an interestingly good stability of results. Similar conclusions were drawn from results for the other two procedures. These findings imply that there is no concern regarding stability of results from simulation to simulation. The $AARD$ values for these multiple synthetic populations are very close to those reported in Table 7.4 for each of the procedures. The Chi-square test for the difference between the real and synthetic joint distributions also reveals no significant differences for each of the procedures at both the household and person-levels. In general, this result is important as it establishes that the procedures yield stable results that also do not deviate from those attained from use of actual data.

Table 7.10: Household-level joint distribution by multiple runs

Groups	Control variables			Synthetic run									
	AutoYesNo	hhsizeRegrp4	AgeGrpOldest	1	2	3	4	5	6	7	8	9	10
1	no	1	18-59	5054	4964	5018	5049	5038	5005	5016	5060	5044	4991
2	no	1	60plus	9518	9505	9508	9507	9504	9507	9514	9526	9508	9505
3	no	2	18-59	1502	1502	1493	1491	1501	1493	1496	1483	1490	1489
4	no	2	60plus	3154	3145	3143	3148	3141	3146	3147	3155	3146	3143
5	no	3	18-59	705	704	706	699	702	700	703	699	707	701
6	no	3	60plus	341	339	341	338	337	338	339	340	340	342
7	no	4+	18-59	641	648	643	644	644	649	642	628	648	643
8	no	4+	60plus	132	134	134	134	134	136	133	134	136	135
9	yes	1	18-59	15783	15579	15747	15797	15761	15651	15701	15804	15783	15636
10	yes	1	60plus	7155	7147	7153	7157	7153	7153	7148	7169	7148	7150
11	yes	2	18-59	19077	19026	18952	19087	19004	19056	19062	19046	18992	19019
12	yes	2	60plus	19094	19062	19077	19081	19091	19072	19082	19119	19095	19066
13	yes	3	18-59	15468	15403	15378	15442	15427	15450	15433	15421	15430	15411
14	yes	3	60plus	3502	3506	3495	3511	3506	3508	3500	3510	3503	3506
15	yes	4+	18-59	22736	22755	22751	22725	22784	22788	22752	22697	22796	22755
16	yes	4+	60plus	1796	1816	1797	1810	1804	1806	1811	1806	1808	1803
Total				125658	125235	125336	125620	125531	125458	125479	125597	125574	125295

Table 7.11: Person-level joint distribution by multiple runs

Groups	Control variables		Synthetic run									
	gender	age grp	1	2	3	4	5	6	7	8	9	10
1	1	1	6909	6644	6648	6814	6759	6850	6629	6796	6858	6748
2	1	2	8311	8311	8311	8311	8311	8311	8311	8311	8311	8313
3	1	3	8949	8839	8881	8735	8949	8785	8769	8734	8868	8833
4	1	4	9234	9234	9234	9234	9234	9234	9234	9100	9234	9234
5	1	5	9135	9135	9135	9135	9136	9136	9136	9135	9135	9135
6	1	6	9716	9716	9716	9716	9716	9716	9716	9716	9716	9716
7	1	7	9657	9657	9657	9657	9657	9657	9657	9657	9657	9657
8	1	8	11177	11177	11177	11177	11177	11177	11177	11177	11177	11177
9	1	9	12355	12355	12355	12355	12355	12355	12355	12355	12355	12355
10	1	10	11994	11994	11994	11994	11994	11994	11994	11994	11994	11994
11	1	11	10908	10908	10908	10908	10908	10908	10908	10908	10908	10908
12	1	12	9836	9836	9836	9836	9836	9836	9836	9836	9836	9836

Continued on next page

continued from previous page

Groups	Control variables		Synthetic run									
	gender	age grp	1	2	3	4	5	6	7	8	9	10
13	1	13	7764	7781	7757	7782	7622	7733	7753	7694	7796	7901
14	1	14	6986	6986	6986	6986	6986	6986	6986	6986	6986	6986
15	1	15	6306	6319	6255	6179	6233	6242	6307	6398	6370	6250
16	1	16	4496	4605	4429	4561	4671	4586	4545	4595	4467	4547
17	1	17	2839	2724	2923	2829	2956	2825	2790	2845	2818	2868
18	1	18	750	755	779	786	772	789	750	797	757	751
19	1	19	260	261	272	265	254	276	282	269	265	266
20	2	1	6193	6028	6054	6132	6143	6142	6187	6125	6215	6162
21	2	2	7961	7961	7961	7961	7961	7961	7961	7913	7956	7961
22	2	3	8431	8452	8382	8506	8289	8479	8385	8506	8368	8506
23	2	4	8385	8441	8553	8423	8510	8466	8496	8432	8333	8481
24	2	5	8801	8801	8802	8801	8801	8801	8801	8801	8801	8801
25	2	6	9411	9411	9411	9411	9411	9411	9411	9411	9411	9411
26	2	7	9363	9363	9363	9363	9363	9363	9363	9363	9363	9363
27	2	8	10780	10780	10780	10780	10780	10780	10780	10780	10780	10780
28	2	9	11930	11930	11930	11930	11930	11930	11930	11930	11930	11930

Continued on next page

continued from previous page

Groups	Control variables		Synthetic run									
	gender	age grp	1	2	3	4	5	6	7	8	9	10
29	2	10	11573	11573	11573	11573	11573	11573	11573	11573	11573	11573
30	2	11	10520	10520	10520	10520	10520	10520	10520	10520	10520	10520
31	2	12	9643	9643	9643	9643	9643	9643	9643	9643	9643	9643
32	2	13	7996	7902	7882	8047	7947	8116	7946	7939	8009	7994
33	2	14	7023	7143	7006	6909	7107	7143	7115	7071	7149	7048
34	2	15	7677	7677	7677	7677	7677	7677	7677	7677	7677	7677
35	2	16	6387	6520	6359	6369	6544	6417	6487	6505	6414	6282
36	2	17	3973	3804	3933	3917	4022	3714	3857	3860	3816	3915
37	2	18	2366	2355	2389	2480	2329	2395	2293	2380	2367	2378
38	2	19	1357	1359	1357	1359	1358	1357	1357	1357	1358	1359
Total			297352	296900	296828	297061	297434	297284	296917	297089	297191	297259

Table 7.12: Person-level joint distribution by multiple runs

Synthetic run	Level	
	Household	Person
1	0.01623020	0.05168003
2	0.01706754	0.05364358
3	0.01714966	0.05178054
4	0.01622045	0.05100351
5	0.01663027	0.05052328
6	0.01606038	0.05035111
7	0.01717996	0.05250644
8	0.01748587	0.05068925
9	0.01456133	0.05206855
10	0.01720153	0.05152641
11	0.01679956	0.05309536
12	0.01528700	0.05022000
13	0.01537083	0.05062197
14	0.01691046	0.05097391
15	0.01703320	0.05473780
16	0.01654716	0.05129456
17	0.01376127	0.05229727
18	0.01731644	0.05382451
19	0.01938390	0.05476117
20	0.01597663	0.05376209
21	0.01775902	0.05022394
22	0.01708903	0.05048019
23	0.01661337	0.04970396
24	0.01703616	0.05327089
25	0.01683295	0.05260752
26	0.01555226	0.05298058
27	0.01724694	0.04784840
28	0.01564477	0.05309137
29	0.01638924	0.05027430
30	0.01742535	0.05184053

7.4 The Synthetic Population of Flanders across Years

The *IPU* population synthesis procedure [see Figure 6.1] is integrated as part of the larger detailed synthetic population generation model designed for Flanders as shown in Figure A.1 in the Appendix. The model comprises of three main components: the *input*, *synthesis* and *output*. In the implementation that is conducted in R (*synthesis*), the *IPF* procedure for estimating the target joint distribution precedes the *IPU* method. Other parts of the model(*input* and *output*) are implemented in the SAS software since it is more powerful in handling large data sets. Therefore, the *input* component of the model, which involves significant data manipulation steps and fitting models, is implemented in SAS. This section of the model has been handled and discussed extensively in the earlier Chapters of this thesis. The *output* component of the model then finally involves outputting the synthetic population. Here, once the households and their constituents have been determined from the *IPU*, synthetic household and person identification numbers are assigned and the data are exported to SAS to the model *output* component so as to generate the complete synthetic population with all corresponding household and person-level characteristics. The data are also linked to a synthetic car file that was generated in Chapter 5 to provide complete synthetic car-level data set for the whole population.

In this Section, we now present the results of the synthetic population of Flanders for the years 2001, 2007 and 2021 following the application of the *IPU* procedure. To generate the synthetic population of Flanders of 2001 and 2007, we control for the same set of variables; That is, *HHDER-AGE*, *HH-AUTO* and *HHSIZE* at the household-level, and *P-GENDER* and *P-AGEGRP* at the person-level. The control variables retain their categories as defined before. For the synthetic population of 2021, we control for a slightly different set of variables, determined by the available data. At the household-level, the control variables include: Province (5 provinces of Flanders), *HHSIZE*(1, 2, 3, 4, 5+) and *HH-AUTO* (yes/no). At the person-level, as before, *P-AGEGRP* (19 groups) and gender (male/female) are controlled for.

Table 7.13 shows the total population units of Flanders for 2001, 2007 and 2021. The results are shown for the true as well as the synthetic population of Flanders. It is evidenced that the *IPU* algorithm is able to very closely replicate the existing population totals. As thus, the average household size is also maintained across the years. Table 7.14 displays the distributions of control variables across the years. Results are shown for the actual data of 2001, 2007 and 2021 as well as the synthetic data for the respective years. Data are not available for the distribution of the householder's age in 2021 and results are thus missing here. It is observed according to the true data, that the population of females will be slightly higher by the year 2021. Similarly, this is also observed based on the synthetic data. As the years go by, the younger age groups are becoming smaller whereas the older ones are growing larger. For instance, taking a look at the age category '34less', the population slowly shrinks from 41.73% to 41.35% in 2007 and then 37.55% in 2021. The '70plus' group increases from 6.4% in 2001 to 7% in 2007 and by 2021 the group is over 9%. The '55-75' age group also increases by about 5.6% from 2001 to 2021. This is the irony of the ageing population and it is also clearly captured in the synthetic populations. Given these results, it is interesting to have an overview of the general projected trends of Belgium. According to ECODATA, Federale overheidsdienst Economie and Energy (2009), the population is projected to increase in general over several future years. Figure 7.2 shows the distribution of the projected population of Flanders and Belgium by gender. Here, it is depicted that the number of women will continue to be slightly higher than that of men by more or less the same margin. The epoch of ageing population has been observed in several areas around the world and its influence on travel behavior has been intensely studied (Schomocker et al., 2005, Bush 2005, Alsnih et al., 2003, Currie and Delbosh 2010, Scott 2009, Hensher 2007, Davey 2007). It is evidenced that the total size of the elderly (65+ years) population is on the increase, which potentially implies an overall pool of potential travelers. It is well known that an ageing population can have impact on several sectors of the economy, including the transport sector. This is why the topic of ageing population is currently receiving a lot of attention in many areas of the world. It is therefore important to understand this aspect as we can hardly talk about the future population without putting the ageing population into perspective. According to the US Census Bureau (Staff, 2008), 'Baby boomer' is a North American-English term to describe a person born between 1946 and 1964. They are currently aged between 45-63 years. In

Canada, 'Baby boomers' are born between 1947 to 1966 whereas in Australia they are born between the years of 1946-1961. In the United Kingdom, there was a sharp post-World War II peak in 1947, when more babies were born than in any year since the post-World War I peak in 1920. There was then a decline, followed by a broader but lower peak in the 1960s. Thus British 'Baby boomers' are younger than their American counterparts and had not risen to such prominence when the term was coined. In Belgium, ageing population is clearly evidenced in the changing age distributions over the years. Figure 7.3 displays the population distribution of Flanders and Belgium by age groups. Apparently, the only clear rising trend over the years is that of the '65plus' group. Other age groups either exhibit a falling or stable trend overtime. The implication of this is that the 65+ years age group is enlarging in size slowly closing the gap between the size of the active population (20-59 years) and the elderly population. This unveils a very important issue of attention. There is a potential need to study the impact of this trend on the economy. It may be of interest to set up scenarios of synthetic populations to study other general population characteristics and travel behavior of future populations.

Table 7.13: The population units of Flanders across years

Population of Flanders	Year	No. of Persons	No. of Households	Mean household size
True				
	2001	5 968 074	2 426 614	2.46
	2007	6 117 440	2 547 686	2.40
	2021	6 211 780	2 748 772	2.26
Synthetic				
	2001	5 966 906	2 426 614	2.46
	2007	6 117 783	2 547 686	2.40
	2021	6 203 617	2 748 772	2.26

Table 7.14: The distributions of the control variables across the synthetic populations

Variables	Actual data				Synthetic data		
	FS	*FHTS'07*	Flemish pop. 2007	Flemish pop. 2021	2001	2007	2021
<u>Person level</u>							
Gender							
Male	49.38	49.02	49.31	49.06	49.31	49.40	49.09
Female	50.62	50.98	50.69	50.94	50.69	50.60	50.91
Age categories							
34less	41.73	36.48	41.35	37.55	41.36	40.39	37.41
35-54	29.94	31.79	29.82	25.53	29.81	29.93	25.55
55-75	21.90	23.61	21.80	27.51	21.84	22.21	27.74
76plus	6.43	8.12	7.02	9.41	6.99	7.47	9.31
<u>Household level</u>							
Age group -Oldest householder							
18to59	63.21	63.88	64.74	-	63.46	64.74	55.83
60plus	36.79	36.12	35.26	-	36.54	35.26	44.17
Household size							
1	27.75	28.00	29.57	32.40	27.79	29.57	32.40
2	33.34	33.55	34.04	37.26	33.33	34.03	37.26
3	17.05	16.08	15.99	13.83	17.03	15.99	13.83
4+	21.86	22.36	20.40	16.50	21.84	20.40	16.50
Possession of at least 1 car							
Yes	80.63	81.79	83.36	88.00	80.25	83.36	88.01
No	19.37	18.21	16.64	12.00	19.75	16.64	11.99

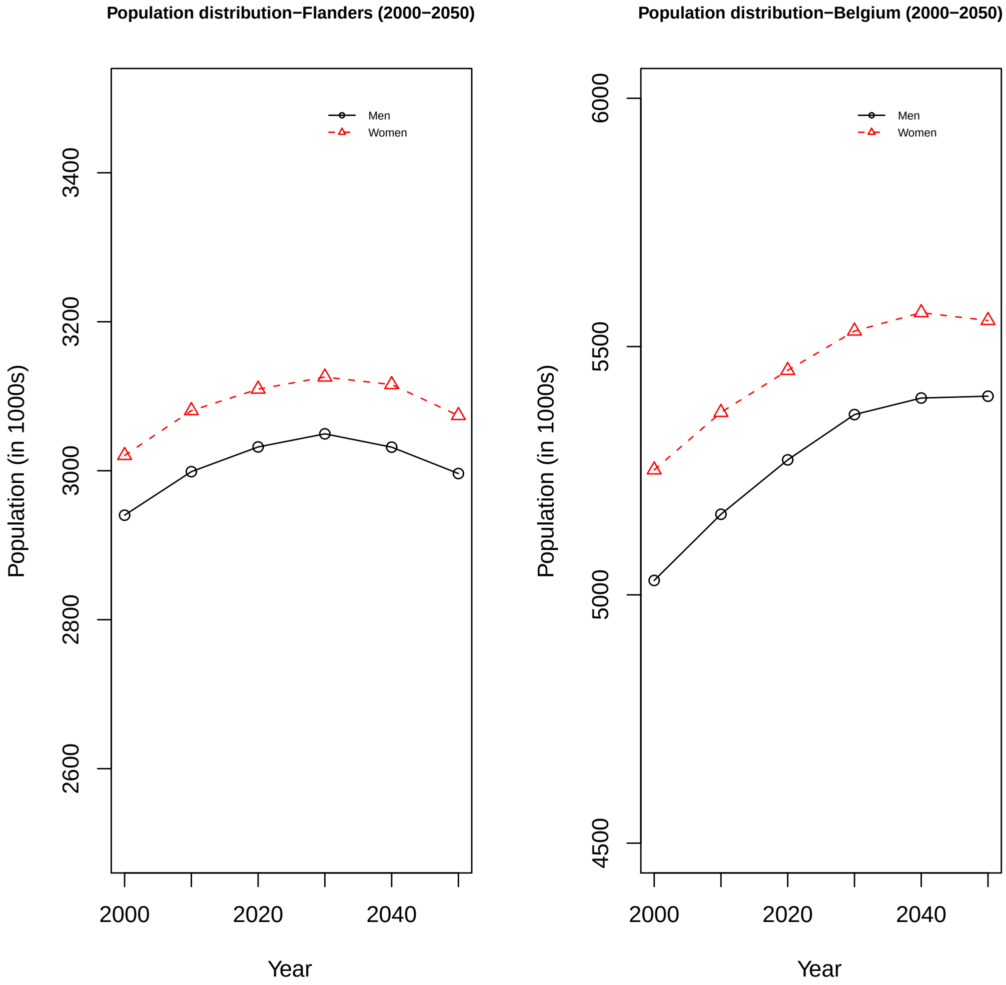


Figure 7.2: The population by gender across years.

In Table 7.14 it can also be observed that the households are also becoming smaller overtime. In 2001 we observe 27.75% single households whereas by 2021 there will be already over 32% single households. The same trend is seen for double households but households of size larger than two depict a reverse trend. Households in possession of at least 1 car are on an increase as well with time. Table 7.15 shows distributions for some variables that were not explicitly controlled for at the household-level. Again the distributions are relatively well maintained in the generated synthetic populations for the

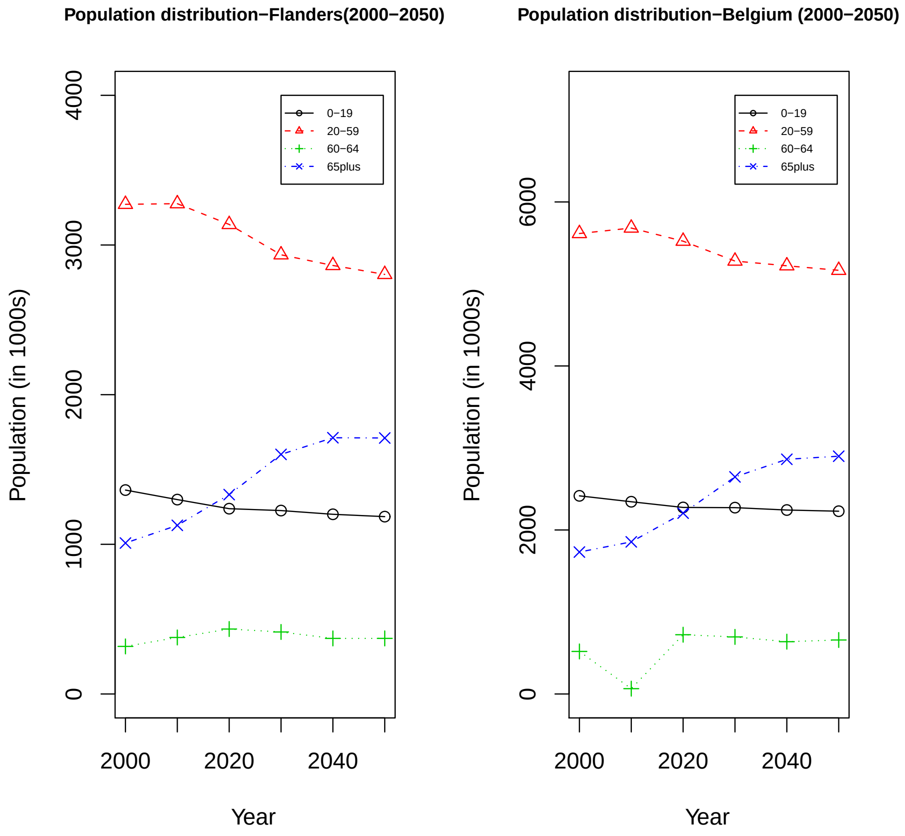


Figure 7.3: The population by age groups across years.

different years. In Table 7.16 some person-level non-control variables are shown. We observed earlier that more households are getting in possession of a car with time. With regard to car usage, more people are also becoming more dependent on cars. It is evidenced that individuals mostly use cars for travel, followed by slow mode of travel and then public transport. Nevertheless a very small increase in public transport usage is also observed. On average, people seem to travel longer for work/school purposes with time. It seems also that the total work hours per week per person will be higher by 2021. This could be good as the working population may need to work much harder and longer to support the ageing population.

Table 7.15: The distributions of some household-level non-control variables

	Actual data				Synthetic data		
Variables	*SEE'01*	*FS*	*FHTS'07*	Flemish pop.2021	2001	2007	2021
No. of cars							
0	19.75	19.37	18.21	-	19.75	16.64	11.99
1	55.36	56.53	53.64	-	56.17	60.16	67.45
2	22.15	21.69	24.75	-	21.66	21.15	18.86
3+	2.74	2.42	3.39	-	2.42	2.05	1.69
No. of bicycles							
0	21.88	22.80	18.56	-	22.77	22.99	24.37
1	24.39	24.23	19.38	-	24.08	25.07	27.04
2	26.81	26.02	25.22	-	26.13	26.42	27.22
3+	26.92	26.96	36.84	-	27.03	25.51	21.36
Province							
Antwerp	28.43	28.40	26.06	28.45	28.48	28.86	28.45
Flemish Brabant	16.98	17.00	17.07	16.78	16.83	17.08	16.78
West Flanders	19.04	19.04	19.09	18.75	19.01	18.72	18.75
East Flanders	23.09	23.06	23.21	22.89	23.24	22.94	22.89
Limburg	12.46	12.50	14.59	13.13	12.45	12.39	13.13

Table 7.16: The distributions of some person-level non-control (unconstrained) variables

Person-level	Actual data			Synthetic data		
variables	*SEE'01*	*FS*	*FHTS'07*	2001	2007	2021
Driver's license						
Yes	64.05	63.84	74.23	64.10	64.47	67.29
No	35.95	36.16	25.77	35.90	35.53	32.71
Personal Income (euros)						
0	24.88	24.99	-	24.89	24.84	22.71
1-1250	21.71	22.54	-	22.17	22.05	20.86
1251-2250	43.12	42.59	-	42.90	43.03	45.50
>2250	10.29	9.88	-	10.03	10.08	10.93
Main mode of work/sch. trips						
Slow mode	22.48	24.16	25.56	24.35	23.74	22.29
Car	61.05	59.88	56.36	59.67	60.05	61.54
Public	13.26	12.93	16.47	12.96	13.18	13.08
Other	3.21	3.02	1.6	3.02	3.04	3.10
Work/School (Y/N)						
Yes	65.87	66.16	66.90	65.67	65.02	59.38
No	34.13	33.84	33.10	34.33	34.98	40.62
Mean(standard dev.)						
Dist. for work/sch. trips	15.18(19.60)	14.50(18.48)	15.95(20.76)	14.50(18.50)	14.62(18.47)	15.07(19.12)
Work hrs per week	39.07(13.19)	38.96(13.21)	-	38.97(13.19)	38.93(13.16)	39.07(13.26)

7.5 Conclusion

Owing to the freshness of the general field of research on synthetic population, several new methods have been recently proposed in literature. As these methods develop, there is lack of comparative studies in which the predictive performance of the competing procedures of population synthesis is compared, more so in a single study. This Chapter has therefore attempted to contribute to the covering of this need. The Chapter has also handled the problem of creating micro-level synthetic data for the population of Flanders over some years; 2001, 2007 and 2021.

The different algorithms implemented in this Chapter have been based on the procedures of population synthesis as proposed by Beckman et al. (1996), Guo and Bhat (2007) and the *IPU* by Ye et al. (2009). The former procedure involves control of only one level of the joint distributions whereas the latter two methods allow simultaneous control for both the household and individual-level distributions. The general population synthesis methodology initially involved use of the Iterative Proportional Fitting algorithm to estimate the household and individual joint distributions of the population for the year 2007. In the next step, the synthesis algorithms were then applied for selection of households for the Flemish synthetic population. Following the algorithms, synthetic data sets were generated as a result of control for the household-level distribution only, as well as simultaneous control for both the household and the person-level joint distributions. The synthetic data sets were then compared across procedures as well as to actual data. Given the general difficulty of simultaneously meeting both the target household and person-level distributions, the possibility of allowing for over-sampling by permitting for a percentage deviation from the target distribution was also investigated.

The pros and cons of each approach have been revealed, based on the applications provided here including both their algorithmic efficiency and the accuracy of the resulting synthetic data. Whereas all implemented procedures consume a reasonably practical amount of computation time to generate a synthetic population, the implemented *IPU* approach provided the most efficient performance and also yielded a perfect match of the control household joint distribution (which is allowed for by design) as well as clearly providing the best fit of the target person-level joint distribution. A strong point of the *IPU* is its easier applicability in practice.

Synthesis procedures have played an important role in providing the required data that are used in *FEATHERS*. Synthetic populations have been constructed for the years 2001, 2007 and 2021. In synthesizing the population for 2001 and 2007, at the household-level, the variables household size, age of the householder and possession of car were controlled for. Simultaneously, the gender and the age of individuals within these households were controlled for at the person-level. The synthetic population of Flanders in 2001 consists of 5,966,906 people, belonging to 2,426,614 households. For the year 2007, there were 6,117,783 synthetic individuals in 2,547,686 households. Projections also permit the future population of Flanders to be visualized now. There were 6,203,617 synthetic individuals created for the population in 2021. These belong to 2,748,772 synthetic households.

Overall, the synthetic procedures controlling for both the household and the person-level variables are able to reasonably preserve the target control household and person-level joint distributions. More still, the distributions of the non-control variables are also estimated impressively well. General travel related variables such as travel distance for work/school trips and working hours per week are estimated quite precisely by all the synthesis procedures. The result on comparison of the distributions of both uncontrolled household and person-level attributes against observed counterparts has been a major contribution to the research field of synthetic populations as all the earlier works only compare controlled attributes. The findings encourages further research on how a good synthetic population improves travel demand forecast. The two algorithms; Guo and Bhat (2007)'s method and the *IPU*) perform well and provide a valuable methodology of population synthesis with none of the two proving to be out-rightly inferior to the other. The results obtained from comparing the generated synthetic population with the real data provided support that both the household and the individual-level distributions of the control and some non-control variables represent the true population rather well and that the actual population could be relatively accurately synthesized. Interestingly also, when an alternative sample is used as opposed to the original one, overall, the findings demonstrate that no major divergencies are observed in the generated synthetic data. This result is important towards a generalization of the results.

There are still some open problems for further research. It may be of interest to set up scenarios of synthetic populations to study various detailed population characteristics in relation to travel behavior of future populations. Within this objective, the role of ageing population should be an important focus. Further studies on creation of synthetic data sets with more diverse case studies can be used to test if the findings can be generalized. Moreover, we acknowledge that there are other issues involved in synthesizing data for populations. For instance representation of individual and/or household heterogeneity based on an unlimited set of variables. This should be an interesting direction of future research. Furthermore, although our research on synthetic populations has achieved the intended goals, there is a limitation that the population is not drawn for small geographic regions. If there is interest in achieving representativeness up to much smaller geographic units than considered here, it would be imperative that distributions are controlled at the respective geographic levels of interest. This objective should indeed be an interesting one if it is supported by availability of all the required data to level of detail necessary to conduct the exercise. Finally, a natural extension of this research in future explorations could be to apply the *FEATHERS* model to the Flemish synthetic population to compare the travel participation shares to those reported in surveys or those generated when *FEATHERS* is applied on actual data.

8 Final Conclusions

Several approaches have recently been proposed by different researchers in attempting to bridge the gap between available data and high data needs. The approaches generally rely on the different forms of the available data to synthetically create further data. Population synthesis frequently involves generation of large amounts of data through procedures that are generally computationally intensive. In the past, limitations - both computational and methodological, have limited the use and development of more sophisticated approaches. Recent developments however, have broadened the possibilities available to researchers. As a consequence, new methodologies to generate synthetic data are being developed and proposed. Most of these new methods have not yet been understood by the research community as they are in their early development stages and policy makers are still reluctant to adopt these methods.

In this thesis, we have attempted to enhance further understanding of these methods through applying the methods to different research problems of focus that required creation or availing new data. A great part of this thesis has thus been devoted to reviewing and assessing different methods under different scenarios. This research has further served to demonstrate, review and/or propose modifications to existing approaches geared towards providing useful and important information regarding their application. More to this, an integrated model has been proposed for generating synthetic populations for Flanders, which was the main goal of this study.

One of the problems that has been handled in this thesis is data integration. This has been dealt with in Chapter 3. Presently, statistical matching constitutes a rich variety of applications in areas including micro-simulations, marketing and official statistics. In this research, statistical matching was conducted enriching the available Flemish household travel survey data with information from a time use survey. The technique enabled adding new data to existing survey data. The matching was effected based on some common variables between a household travel survey data set and a time use survey file. The results from comparison of the resultant synthetic file to the original data demonstrated that statistical matching can yield results that are substantially comparable with actual data. Basic statistics such as the mean and standard deviations are considerably preserved. Based on the findings, it seems promising to use statistical matching as a tool for integration, to supplement travel survey data. This practical experience also revealed some difficulties relating to data preparation, with respect to which guidelines have been proposed. Given that the relation of variables is examined here based on only a single experiment, the generalization of the results are still open to further research. A remedy to this shortcoming may be to apply a simulation-like the bootstrap. The procedure may then involve generating several bootstrap donor files from the original donor file before utilizing statistical matching. It is anticipated that this approach may however involve a heavy computational burden. Future research may also involve integrating information from multiple files using statistical matching while incorporating external information that is not currently available. This will permit a more effective match and validation of the assumption of conditional independence.

Another challenge in this thesis was the scenario of simulating synthetic individual-level disaggregate travel data that was detailed in Chapter 4. A simulation approach based on Stopher et al. (2003) was employed for this task. This approach was developed further by proposing an alternative grouping technique and also expanding the method to incorporate an in-depth validation approach. This new approach was tested on the case study of Flanders and results were compared with those obtained from the former approach. In the application, interest was particularly focused on the simulation of trip rates and duration data. A motivating case study highlighted on the issue of precision from simulation approaches. Results from the modeling approach where the Poisson regression model was fitted to both actual and simulated data were presented. It is particularly interesting to note that parameter estimates attained from simulated data were not generally different from those from actual data. Despite the few observed differences, overall, the parameter estimates were found to be roughly consistent in direction and magnitude, with similar estimates of standard errors. The ranges of the confidence intervals for the respective parameters were also observed to be relatively similar. The results of this application revealed an interesting finding that simulated travel data can be replicative of actual travel data.

In an endeavor to avail some particular data of interest in this study, different approaches were then proposed in Chapter 5. Time series models allowed for prediction of car ownership for Belgium up to the year 2021. While a good number research initiatives on car ownership rely a lot on discrete choice models, the setting of the data required in this study was rather different from these studies and this therefore called for proposition of different models. Overall, the models applied for the current problem sought to isolate trends from irregular variation and to incorporate the found patterns into the forecast. The Holt's linear exponential smoothing model and the Autoregressive Moving Average ($ARMA$) modeling approaches attempted to explain current and future values of each response variable as a weighted average of the variable's own past values. A Box-Tiao model, which corrects for autocorrelation by describing the errors terms of the linear regression model by an $ARMA$ process was also investigated. Furthermore, in addition to taking care of the past values of the response series and past errors, the Box-Tiao method offers the opportunity to model the response series using the current and past values of other series that may be useful in explaining the variability in the response. The Box-Tiao model was ultimately proposed for prediction of car ownership as the analyses on model comparison and validation revealed that this model was superior to the other models, providing more accurate predictions over the forecast period. The results generally revealed that car ownership demand will continue to increase over the next decade in Belgium. With a public policy perspective, the results of this research are of interest. Towards sustainable transportation, transport planners in Belgium are therefore encountered with a challenge of devising countermeasures capable of effectively curtailing ownership and probably, usage of high-emissions and low fuel efficiency cars. The impact of increasing motorization and infrastructure expansion on the environment, health, energy demand and land-use may consequently require attention. Another interesting issue is that, one may not completely rule out the notion that the demand for cars can become saturated. The model proposed here has not revealed such patterns over the forecast period. In future, it may be of interest to develop a model of the relationship between economic development and per capita private car ownership. Such research may also focus on investigating the factors that influence the evolution of vehicle stocks. Furthermore, if the goal is devising effective management strategies to relieve dependency on private vehicles, *i.e.* cars and motorcycles, then disaggregate choice models regarding the ownership, type and usage of cars and motorcycles may be required to achieve this. It would be interesting to model choice behaviors related not only to ownership and usage, but also to car type, since gas mileage and emission coefficient of cars differ considerably with engine size and age. It could also be of interest to investigate if energy consumption and emissions vary markedly across different engine sizes and ages.

The necessity to create a micro-level car file for use in *FEATHERS* also geared the proposal of a car mileage model in Chapter 5. The model followed a simulation-based approach through which variables that were of interest were simulated. The variables that were contained in the created car file include car mileage, fuel type and age of a car. In future, other variables such as car type would also be of interest to include in this file when the available data permit their creation. It should be interesting also to investigate other approaches for instance, those that would make use of multi-level models to achieve a similar objective. Such methods would account for possible correlation within clusters that is possibly not explicitly taken care of by the proposed model. The concern that would motivate such research would be that, car units on which the data are observed can be grouped into clusters; the households, and the data from a common cluster are possibly correlated. A potentially interesting approach to investigate in future, could be one that utilizes a linear mixed modeling framework for vehicle mileage.

Further micro-level research objectives were taken on and solutions were proposed in this thesis. A population census file was available for Belgium but it lacked some important variables that were needed within *FEATHERS*. A Proportional odds model was thus proposed in Chapter 5 for prediction of personal income and a Logistic regression model was also proposed to predict possession of driver's licence. These models were very suited for these problems given the nature of these variables and they performed reasonably well.

The challenge of laying down the framework for creating synthetic populations was the main topic in Chapter 6. Several methods were of interest in creating synthetic populations but attention was focused on three methods that were investigated for the ultimate purpose of generating synthetic populations. These methods include: the method proposed by Beckman et al. (1996), the second approach by Guo and Bhat (2007) and the *IPU* algorithm that was proposed by Ye et al. (2009). Population synthesis being a new field of research, most of the methods are fresh proposals that can benefit from a critical and evaluative eye. The methods were conceptually evaluated to provide more enlightenment and the formal mathematical formulations which have been lacking were also proposed. The ideas thus developed in this thesis were presented to feature and provide further insight on the currently existing methodology for the creation of synthetic data.

An integrated model has been developed for generating synthetic populations for Flanders. The model comprises of three main components: the *input*, *synthesis* and the *output*. The *IPU* method has been integrated as part of the *synthesis* component together with the *IPF* method. The *input* component has been designed for in-depth data manipulations and preparation. This component also entails model fitting. The *output* component finally deals with generating the complete synthetic population for Flanders.

In Chapter 7, the problem of creating micro-level synthetic data for the population of Flanders for the years; 2001, 2007 and 2021 was handled. The problem of creating synthetic populations is a well understood problem when units are defined on a single level, either household or persons. When a population needs to be synthesized on both household and person-level at the same time, the problem is less well understood. In this Chapter, a more detailed focus, with an application perspective was provided on this topic. The methods of focus in Chapter 6 were implemented and applied here, and the results have been compared to highlight strength and/or weaknesses of the methodologies and the resultant quality of the synthetic data. The *IPU* was noted to require less computation time as compared to Guo and Bhat (2007), which is a strong point of the former method with regard to applicability in practice. The *IPU* algorithm proved to be particularly useful as it provides an approach that precludes the often tedious and computationally demanding aspects of population synthesizers in generating synthetic data. The research has highlighted that some emerging methods are capable of creating *good* synthetic populations. It has further been established that controlling both household and person-level attributes preserves target control household and person-level joint distributions. Moreover, the distributions of uncontrolled attributes are also reasonably well preserved. There is also stability of results from different synthetic populations generated using these methods. The result on comparison of the distributions of both uncontrolled household and person-level attributes against observed counterparts has been a major contribution to the research field on generation of synthetic populations as all the earlier works only compare controlled attributes. The findings encourage further research on how a ‘good’ synthetic population improves travel demand forecast. There are still some open problems for further research. It may be of interest to set up scenarios of synthetic populations to study various detailed population characteristics in relation to travel behavior of future populations. Within this objective, the role of ageing population should be an important focus. Further studies on creation of synthetic data sets with more diverse case studies can be used to test if the findings can be generalized. Moreover, is is acknowledged that there are other issues involved in synthesizing data for populations. For instance representation of individual and/or household heterogeneity based on an unlimited set of variables. This should be an interesting direction of future research. Furthermore, although our research on synthetic populations has achieved the intended goals, there is a limitation that the population is not drawn for small geographic regions. If there is interest in achieving representativeness up to much smaller geographic units than considered here, it would be imperative that distributions are controlled at the respective geographic levels of interest. This objective should indeed be an interesting one if it is supported by availability of all the required data to level of detail necessary to conduct the exercise. Finally, a natural extension of this research in future explorations could be to apply the *FEATHERS* model to the Flemish synthetic population to compare the travel participation shares to those reported in surveys or those generated when *FEATHERS* is applied on actual data.

The choice regarding the type of methods to be employed is usually dictated by several pragmatic considerations, which typically include the particular scientific objectives of interest, computational issues, as well as the magnitude and nature of data required to be created. Because such a choice can often be daunting to the researcher, not to mention the practitioner, the examination of the different procedures considered herein has hopefully provided additional useful information regarding their use. Alongside highlighting the relative merits of various techniques available to the researcher, the thesis has also solved practical problems including creation of synthetic populations for Flanders. These data are being used for micro-simulation in an operational model *FEATHERS* (Forecasting of Evolutionary Activity-Travel of Households and their Environmental RepercussionS), which is vital for guiding and supporting transportation policy in Flanders and Belgium in general.

Appendices

A Population synthesis: Implementation

The main parts of the procedures for generating synthetic populations are implemented in the R language. A guide to these implementations is given below in form of Pseudocode. R is a statistical computing platform whose syntax closely resembles S-PLUS, but with an underlying implementation borrowed from the Scheme and Lisp languages. It was selected largely because of good performance. While it was suitable for prototyping and experimenting with new methods, its data storage is not efficient for large amounts of data, and its performance is poorer than low-level languages like C. Therefore most data manipulations were performed in SAS version 9.1.

The implementation of the algorithm proposed by Guo and Bhat (2007) largely followed the description and formal mathematical formulation we proposed in Chapter 6. Its inputs include a data frame for the estimated joint distribution at the household-level to be referred to as the *HH* and that for the individual-level *IND*. The corresponding zero-initialized files, are then referred to as *HHU* and *INDU* respectively. The details of the implementation then follow the pseudocode shown in Table A.1.

The implementation of the *IPU* algorithm proposed by Ye et al. (2009) also largely followed the description and formal mathematical formulation proposed in Chapter 6. Its inputs include a data matrix $\mathbf{X}_{N_h*M}$ derived from the base sample (*FS*). The matrix $\mathbf{X}_{N_h*M}$ is characterized by special input format that represents counts at the household-level with respect to the joint distribution groups for both household and person-level variables [see Ye et al. (2009) for an example]. A data vector $\mathbf{V}_r$ for estimated joint distribution at the household-level together with the person-level is also required. The procedure is then implemented following the pseudocode shown in Table A.2. The implementation of the algorithm proposed by Beckman et al. (1996) is quite straightforward and is not identified here.

Table A.1: The steps in implementing the procedure by Guo and Bhat (2007)

```
1. Initialization
 ⇒ Call input files $HH$, $IND$ and FlemishSample
 ⇒ Create initialized files $HHU$ and $INDU$ for tracking
    the distributions of selected households and individuals
 ⇒ Initialize all desired matrices and constants such as the threshold,
       stop criteria and so forth.

WHILE(stop criteria not satisfied for threshold(s))
{

  2. Calculate new selection probabilities for each household with respect to
    each demographic group

  3. Household selection
   ⇒Draw 1 household($H_{l^\star}^{(i^\star)}$) from the sample - FlemishSample
   ⇒Retrieve the person(s) $P_{l^\star,1}^{j_1^\star}, P_{l^\star,2}^{j_2^\star}, \ldots, P_{l^\star,n_{l^\star}}^{j^\star_{n_{l^\star}}}$ belonging to household

  4. Evaluate the desirability of the household
    IF target distribution $HHU$ for household-type $i$ is exceeded
         ⇒exclude $H_{l^\star}^{(i^\star)}$ from the sample-$FS$
         ⇒update files required for re-calculation of selection probabilities accordingly

    IF ($HHU$ distribution not satisfied or just met group $i$)
         IF any person-type distribution (to which any member of $l$ belongs) is met
              ⇒exclude ($H_{l^\star}^{(i^\star)}$) from the sample-$FS$
              ⇒update files required for re-calculation of selection
                probabilities accordingly
         END IF-loop

         IF NO person-type distribution (to which all persons from $l$ belong) is met
          ⇒Update $HHU$ by 1
              FOR each person in this selected $l$
                   ⇒ Update $INDU$ for all person-types $j^\star$
              END FOR-loop
              Add household $H_{l^\star}^{(i^\star)}$ and constituent persons to the synthetic population
         END IF-loop
    END IF-loop
    ⇒update threshold
}
```

Table A.2: The steps in implementing the *IPU* procedure

```
Part I:Initialization
 -Initialize weights $w_l$  -Initialize $\theta_{min}$
 -Initialize $\theta_{diff}$

Part II: Calibrate IPU weights
WHILE(stop criteria not satisfied for threshold)
{
        Initialize parameters
        FOR loop:over $M$
                        -Compute adjustment factor $\phi$
                        -Update weights $w^{\star}$
        END FOR-loop
        -Update $\theta^{\star}$
             FOR loop:over $M$
                               -Compute $\theta_s$
                               -Compute current $\theta$
             END FOR-loop
             -Update $\theta_{diff}$
             IF(current $\theta < \theta_{min}$)
                               -Update $\theta_{min}$
                               -Update weights
             END IF-loop
  Iterate until convergence
}

Part III: Mapping the household distribution perfectly
{
    Initialize weights with weights from previous step
      $M_h$ =to household-level types
             FOR loop:over $M_h$
                               -Compute adjustment factor $\phi$
                               -Update weights $w^{\star}$
             END FOR-loop
    =>Obtain Final $IPU$ weights
}
```

A.1 Detailed synthetic population generation model

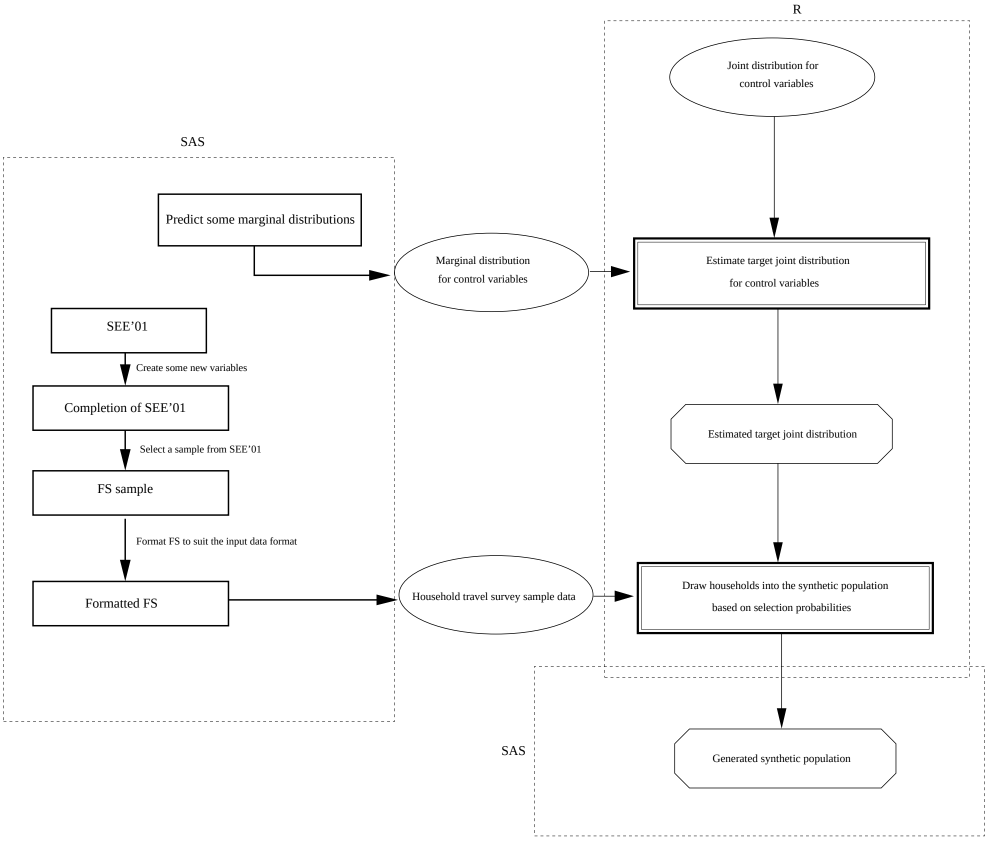

R
Joint distribution for control variables
SAS
Predict some marginal distributions
Marginal distribution for control variables
Estimate target joint distribution for control variables
SEE'01
Create some new variables
Completion of SEE'01
Select a sample from SEE'01
FS sample
Estimated target joint distribution
Format FS to suit the input data format
Formatted FS
Household travel survey sample data
Draw households into the synthetic population based on selection probabilities
SAS
Generated synthetic population

Figure A.1: Detailed model for generating synthetic populations.

B Data

Table B.1: Household-level joint distribution for the Flemish population in 2001 - $SEE'01$

HH-AUTO	*HHSIZE*	*HHDER-AGE*	*COUNT*	*HH-AUTO*	*HHSIZE*	*HHDER-AGE*	*COUNT*
no	1	18-59	104784	yes	1	18-59	243673
no	1	60plus	208935	yes	1	60plus	117063
no	2	18-59	35966	yes	2	18-59	339651
no	2	60plus	78607	yes	2	60plus	354584
no	3	18-59	19177	yes	3	18-59	310709
no	3	60plus	9786	yes	3	60plus	73628
no	4	18-59	10849	yes	4	18-59	323720
no	4	60plus	2125	yes	4	60plus	21575
no	5	18-59	4280	yes	5	18-59	107521
no	5	60plus	820	yes	5	60plus	9571
no	6	18-59	1539	yes	6	18-59	26424
no	6	60plus	404	yes	6	60plus	4571
no	7	18-59	621	yes	7	18-59	6253
no	7	60plus	181	yes	7	60plus	1791
no	8	18-59	259	yes	8	18-59	2280
no	8	60plus	98	yes	8	60plus	746
no	9	18-59	121	yes	9	18-59	965
no	9	60plus	45	yes	9	60plus	332
no	10+	18-59	196	yes	10+	18-59	979
no	10+	60plus	393	yes	10+	60plus	1392
Total							2426614

Table B.2: Personal-level Joint distribution for the Flemish population in 2001 - $SEE'01$

P-GENDER	*P-AGEGRP*	*COUNT*	*P-GENDER*	*P-AGEGRP*	*COUNT*
male	0-4	141604	female	0-4	135769
male	5-9	172047	female	5-9	165210
male	10-14	181006	female	10-14	172481
male	15-19	175553	female	15-19	167598
male	20-24	188308	female	20-24	181873
male	25-29	184274	female	25-29	178949
male	30-34	214722	female	30-34	208723
male	35-39	241962	female	35-39	233966
male	40-44	239674	female	40-44	232016
male	45-49	220535	female	45-49	213337
male	50-54	202345	female	50-54	195653
male	55-59	179940	female	55-59	176850
male	60-64	150381	female	60-64	155543
male	65-69	146660	female	65-69	160206
male	70-74	128415	female	70-74	154480
male	75-79	94614	female	75-79	130389
male	80-84	49547	female	80-84	82415
male	85-89	21852	female	85-89	49416
male	90plus	9261	female	90plus	30500
Total					5968074

Table B.3: Household-level marginal values for the householders aged 18 and above - Flanders 2007

HHDERAGE		*HH-AUTO**		*HHSIZE*	
categories	marginals	categories	marginals	categories	marginals
18-59	1649346	Yes	2122222	1	753355
60plus	898340	No	425464	2	867108
				3	407388
				4	351938
				5	119845
				6	32864
				7	9149
				8	3367
				9	1359
				10+	1313
Totals	2547686		2547686		2547686

*Values obtained from prediction models [See Chapter 5]

Source: '*Studiedienst van de Vlaamse Regering (2009)*'

Table B.4: Person-level marginal values for the householders aged 18 and above - Flanders 2007

P-GENDER		*P-AGEGRP*	
categories	marginals	categories	marginals
male	3017063	0-4	319246
female	3100377	5-9	325444
		10-14	349117
		15-19	360564
		20-24	358716
		25-29	382534
		30-34	380393
		35-39	439136
		40-44	485714
		45-49	471326
		50-54	428560
		55-59	389589
		60-64	337794
		65-69	291961
		70-74	281509
		75-79	238323
		80-84	167003
		85-89	75099
		90plus	35412
Totals	6117440		6117440

Source: ‘*Studiedienst van de Vlaamse Regering (2009)*’

Table B.5: Household-level marginal values for the householders of Flanders 2021

PROVINCE		*HHSIZE*		*HH-AUTO**	
categories	marginals	categories	marginals	categories	marginals
Antwerp	781972	1	890560	Yes	2419046
Vlaams Brabant	461280	2	1024284	No	329726
West Flanders	515355	3	380288		
East Flanders	629209	4	308234		
Limburg	360956	5+	145406		
Totals	2748772		2748772		2748772

*Values obtained from prediction models [See Chapter 5]

Source: '*Studiedienst van de Vlaamse Regering (2009)*'

Table B.6: Person-level marginal values for Flanders in 2021

P-PROVINCE		*P-GENDER*		*P-AGEGRP*	
categories	marginals	categories	marginals	categories	marginals
Antwerp	1744638	male	3047754	0-4	311705
Vlaams Brabant	1071183	female	3164026	5-9	317916
West Flanders	1138977			10-14	317865
East Flanders	1402752			15-19	315677
Limburg	854230			20-24	336967
				25-29	364578
				30-34	367719
				35-39	373653
				40-44	383664
				45-49	389073
				50-54	439316
				55-59	474687
				60-64	443357
				65-69	389086
				70-74	343756
				75-79	250202
				80-84	197159
				85-89	128444
				90plus	66956
Totals	6211780		6211780		6211780

Source: '*Studiedienst van de Vlaamse Regering (2009)*'

Bibliography

Agresti, A. (2002). *Categorical data analysis* (2 ed.). Hoboken, New Jersey: John Wiley and Sons Inc.

Alter, H. (1974). Creation of a synthetic data set by linking records of canadian survey of consumer finances with family expenditure survey 1970. *Annals of economic and social measurement 3*(2), 373–394.

Aluja-Banet, T. and S. Thio (2001). Survey data fusion. *Bulletin of Sociological Methodology* (72), 20–36.

Angrist, J. and A. Krueger (1992). The effect of age at school entry on educational-attainment - an application of instrumental variables with moments from 2 samples. *Journal of the American Statistical Association 87*(418), 328–336.

Arellano, M. and C. Meghir (1992). Female labor supply and on-the-job search - an empirical-model estimated using complementary data sets. *Review of Economic Studies 59*(3), 537–557.

Arentze, T., F. Hofman, H. van Mourik, and H. Timmermans (2000). Albatross: Multiagent, rule-based model of activity pattern decisions. *Transportation Research Record: Journal of the Transportation Research Board 1706*, 136–144.

Arentze, T. and H. Timmermans (2005). *Albatross version 2: A learning-Based Transportation Oriented Simulation System.* Eindhoven, The Netherlands: European Institute of Retailing and Services Studies.

Arentze, T., H. Timmermans, and F. Hofman (2007). Creating synthetic household populations: Problems and approach. *Transportation Research Record: Journal of the Transportation Research Board 2014*, 85–91.

Auld, J. and A. K. Mohammadian (2009). An efficient methodology 1 for generating synthetic populations with multiple control levels. Washington D.C.

Auld, J. A., A. K. Mohammadian, and K. Wies (2009). Population synthesis with Subregion-Level control variable aggregation. *Journal of Transportation Engineering 135*(9), 632.

Barcena, M. J. and F. Tussel (2000). Multivariate data imputation using trees. In *COMPSTAT Proceedings in Computational Statistics*, pp. 193–204.

Beckman, R. J., K. A. Baggerly, and M. D. McKay (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice 30*(6), 415–429.

Belgian Federal Government (2009a). Eonomie. http://ecodata.mineco.fgov.be/.

Belgian Federal Government (2009b). http://www.statbel.fgov.be/figures/d24_nl.asp. http://www.statbel.fgov.be/figures/d24_nl.asp.

Bellemans, T., D. Janssens, G. Wets, T. Arentze, and H. Timmermans (2010). Implementation framework and development trajectory of the FEATHERS Activity-Based simulation platform. *Transportation Research Record: Journal of the Transportation Research Board, In press.*

Bhat, C., J. Guo, S. Srinivasan, and A. Sivakumar (2004). Comprehensive econometric microsimulator for daily activity-travel patterns. *Transportation Research Record: Journal of the Transportation Research Board 1894*, 57–66.

Bhat, C. R. and V. Pulugurta (1998). A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological 32*(1), 61–75.

Birkin, M. and M. Clarke (1988). SYNTHESIS – a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A 20*(12), 1645–1671.

Bishop, Y. M. M., S. E. Fienberg, P. W. Holl, R. J. Light, F. Mosteller, and P. B. Imrey (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: The MIT Press.

Boman, M. and E. Holm (2004). Multi-agent systems, time geography, and microsimulations. In O. Olsson and G. Sjöstedt (Eds.), *Systems Approaches and their Application*, pp. 95118. Kluwer Academic.

Boman, M. and S. J. Johansson (2007). Modeling epidemic spread in synthetic populations - virtual plagues in massively multiplayer online games . *The Computing Research Repository*.

Bowman, J. and G. Rousseau (2006). Validation of the atlanta (ARC) population synthesizer (PopSyn). Austin, Texas.

Box, G. and G. Jenkins (1976). *Time Series Analysis Forecasting and Control.* San Francisco, Holden Day.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984). *Classification and Regression Trees* (1 ed.). Chapman & Hall.

Brocklebank, J. C. and D. A. Dickey (2003). *SAS for Forecasting Time Series* (2 ed.). WA (Wiley-SAS).

Brouwers, L. (2005). Micropox: a large-scale and spatially explicit microsimulation model for smallpox planning. In *proceedings Western Simulation Multiconf, Intl Conf Health Sciences Simulation*, New Orleans. The Society for Modeling and Simulation International.

CBS (2009). Central Bureau of Statistics: CBS StatLine. http://statline.cbs.nl/statweb/.

Chiou, Y., C. Wen, S. Tsai, and W. Wang (2009). Integrated modeling of car/motorcycle ownership, type and usage for estimating energy consumption and emissions. *Transportation Research Part A: Policy and Practice 43*(7), 665–684.

Chipman, H. and R. E. McCulloch (2000). Hierarchical priors for bayesian CART shrinkage. *Statistics and Computing 10*(1), 17–24.

Clarke, G. P. (1996). *Microsimulation for urban and regional policy analysis.* Pion Ltd.

Cohen, M. L. (1991). Statistical matching and microsimulation models. In C. F. Citro and E. A. Hanushek (Eds.), *Improving Information for Social Policy Decisions: The Uses of MicrosimulationModeling*, Volume 2: Technical Papers. Washington, DC: National Academy Press.

Consult, P. (2005). MORPC travel demand model, validation and final report. Technical report, Mid-Ohio Regional Planning Commission, Columbus, Ohio.

Conversano, C. and R. Siciliano (2002). Tree based classiffiers for conditional incremental missing data imputation. In *Proceedings of Data Clean 2002 Conference.*

Cools, M., E. Moons, T. Bellemans, D. Janssens, and W. Geert (2009). Surveying activity-travel behavior in flanders: assessing the impact of the survey design. In C. Macharis and L. Turcksin (Eds.), *Proceedings of the BIVEC-GIBET Transport Research Day 2009, Part II*, pp. 727–741. Brussels, Belgium: VUBPRESS.

Cools, M., E. Moons, and G. Wets (2007). Investigating effect of holidays on daily traffic counts: Time series approach. *Transportation Research Record: Journal of the Transportation Research Board 2019*(-1), 22–31.

Dargay, J. and D. Gately (1999). Income's effect on car and vehicle ownership, worldwide: 1960-2015. *Transportation Research Part A: Policy and Practice 33*(2), 101–138.

Dargay, J. and P. Vythoulkas (1999). Estimation of a dynamic car ownership model: a pseudo-panel approach. *Journal of Transport Economics and Policy 33*(3), 287–302.

Dargay, J. M. (2001). The effect of income on car ownership: evidence of asymmetry. *Transportation Research Part A: Policy and Practice 35*(9), 807–821.

Davidson, W., R. Donnelly, P. Vovsha, J. Freedman, S. Ruegg, J. Hicks, J. Castiglione, and R. Picado (2007). Synthesis of first practices and operational research approaches in activity-based travel demand modeling. *Transportation Research Part A: Policy and Practice 41*(5), 464–488.

De Jong, M. and J. A. Annema (2010). De geschiedenis van de toekomst: Verkeer- en vervoerscenario's geanalyseerd. Technical Report ISBN 978-90-8902-057-4, KiM 1-A02, Kennisinstituut voor Mobiliteitsbeleid (KiM), Den Haag, Netherlands.

Deming, E. W. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics 11*(4), 427–444.

den Bossche, F. V., G. Wets, and T. Brijs (2004). A regression model with ARMA errors to investigate the frequency and severity of road traffic accidents. In *Proceedings of the 83rd Annual Meeting of the Transportation Research Board*, Washington, D.C.

Denk, M. and P. Hackl (2003). Data integration and record matching: An austrian contribution to research in official statistics. *Austrian Journal of Statistics 32*(4), 305–321.

Denk, M. and P. Hackl (2004). Data integration: Techniques and evaluation. *Austrian Journal of Statistics 33*(1&2), 135–152.

D'Orazio, M., M. D. Zio, and M. Scanu (2006). *Statistical Matching: Theory and Practice* (1 ed.). Wiley.

ECODATA, Federale overheidsdienst Economie, K. M. and Energy (2009). http://ecodata.mineco.fgov.be/Nl/beginnl.htm.

Ettema, D. (1996). *Activity-based travel demand modeling.* Ph. D. thesis, Technische Universiteit Eindhoven, Eindhoven, The Netherlands.

Federal Committee on Statistical Methodology, F. . (1980). Report on exact and statistical matching techniques. Statistical Policy Working paper 5, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, Washington, DC.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics 41*(3), 907–917.

Fienberg, S. E. (1977). *The analysis of cross-classified categorical data* (2 ed.). MIT Press.

Franconi, L. and S. Polettini (2007). Some experiences at istat on data simulation. In *Proceedings of the 56th Session of the International Statistical Institute*, Lisboa, Portugal. ISI.

Gavin, N. I. (1985). An application of statistical matching with the survey of income and education and the 1976 health interview survey. *Health Services Research 20*(2), 183–198.

Gerike, R. (2007). *How to make sustainable transportation a reality.* Library MARC record. (Mnchen): Oekom. ISBN 13: 2008380521.

Giuliano, G. and J. Dargay (2006). Car ownership, travel and land use: a comparison of the US and great britain. *Transportation Research Part A: Policy and Practice 40*(2), 106–124.

Glorieux, I., S. Koelet, and M. Moens (2000). Technisch verslag bij de tijdsbudgetenqute TOR 99.Veldwerk en responsanalyse. Technical report, Department of Sociology, Research Group TOR, VUB, Brussels, Belgium.

Golob, J. and T. F. Golob (1983). Classification of approaches to travel-behavior analysis. Special Report 201, Transportation Research Board, Washington, DC.

Golob, T. F. and L. Burns (1978). Effects of transportation service on automobile ownership in an urban area. *Transportation Research Record: Journal of the Transportation Research Board* (673), 137–145.

Greaves, S. and P. Stopher (2000). Creating a synthetic household travel and activity survey - rationale and feasibility analysis. *Transportation Planning, Public Participation, and Telecommuting* - (1706), 82–91.

Guo, J. and C. Bhat (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board 2014*, 92–101.

Hägerstrand, T. (1970). What about people in regional science? *Papers in Regional Science 24*(1), 6–21.

Hensher, D., P. Stopher, P. Bullock, and T. Ton (2004). TRESIS: application of transport and environmental strategic impact simulator to sydney, australia. *Transportation Research Record: Journal of the Transportation Research Board 1898*, 114–123.

Henson, K., K. Goulias, and R. Golledge (2009). An assessment of activity-based modeling and simulation for applications in operational studies, disaster preparedness, and homeland security. *Transportation Letters: The International Journal of Transportation Research 1*(1), 19–39.

Hunt, J., J. Abraham, and T. Weidner (2004). Household allocation module of oregon2 model. *Travel Demand and Land Use 2004* (1898), 98–107.

Ingram, D. D., J. O'Hare, F. Scheuren, and J. Turek (2000). Statistical matching: A new validation case study. In *Survey Research Methods Section*, pp. 746–751. American Statistical Association.

Ireland, C. T. and S. Kullback (1968). Contingency tables with given marginals. *Biometrika 55*(1), 179–188.

Janssens, D., G. Wets, E. D. D. Beuckeleer, and K. Vanhoof (2004). Collecting activity-travel diary data by means of a new computer-assisted data collection tool. Technical report, Limburgs Universitair Centrum, Diepenbeek.

Jong, G. C. D. (1990). An indirect utility model of car ownership and private car use. *European Economic Review 34*(5), 971–985.

Jong, G. D., J. Fox, A. Daly, M. Pieters, and R. Smit (2004). Comparison of car ownership models. *Transport Reviews: A Transnational Transdisciplinary Journal 24*(4), 379.

Jonnalagadda, N., J. Freedman, W. Davidson, and J. Hunt (2001). Development of microsimulation activity-based model for san francisco: Destination and mode choice models. *Transportation Research Record: Journal of the Transportation Research Board 1777*, 25–35.

Kain, J. and M. Beesley (1965). Forecasting car ownership and use. *Urban Studies 2*, 163–185.

Keister, L. A. (2000). *Wealth in America: Trends in Wealth Inequality* (1 ed.). Cambridge University Press.

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation 15*(1-2), 9–34.

Kitamura, R., E. Pas, C. Lula, T. Lawton, and P. Benson (1996). The sequenced activity mobility simulator (sams): an integrated approach to modeling transportation, land use and air quality. *Transportation 23*(3), 267–291.

Kochan, B., T. Bellemans, D. Janssens, G. Wets, and H. J. P. Timmermans (2010). Quality assessment of location data obtained by the GPS-enabled PARROTS - survey tool. *Journal of Location Based Services 4*(2), 93.

Koelet, S. and I. Glorieux (2007). Finaliteit van de verplaatsingen in de onderzoeken TOR'99, NIS'99 en OVG'00. subnota 2 van het takenpakket 2.3 Synthetic datasets in het kader van het sbo-onderzoek: An activity-based approach for surveying and modelling travel behaviour. Final report, VUB, Brussels.

Kum, H. and T. Masterson (2008). Statistical matching using propensity scores: Theory and application to the levy institute measure of economic Well-Being. Working Paper No. 535, The Levy Economics Institute of Bard College.

LANL, L. A. N. L. (2003). Population synthesizer(chapter 2). In *TRANSIMS-Version 3.0*, Volume 3.

Little, R. and D. Rubin (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health 21*, 121–145.

Little, R. J. A. and M. Wu (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association 86*(413), 87–95.

Lusardi, A. (1996). Permanent income, current income, and consumption: Evidence from two panel data sets. *Journal of Business and Economic Statistics 14*(1), 81–90.

Makridakis, S. G., S. C. Wheelwright, and R. J. Hyndman (1997). *Forecasting: Methods and Applications* (3 ed.). Wiley.

Matas, A. and J. Raymond (2008). Changes in the structure of car ownership in spain. *Transportation Research Part A: Policy and Practice 42*(1), 187–202.

Matas, A., J. Raymond, and J. Roig (2009). Car ownership and access to jobs in spain. *Transportation Research Part A: Policy and Practice 43*(6), 607–617.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models*. CRC Press.

McNally, M. G. (2000). The Activity-Based approach. Technical Report UCIITS-AS-WP-00-4, University of California, Center for Activity Systems Analysis., UC Irvine, USA.

Miller, E. (2003). *Transportation Systems Planning: Methods and Applications (Hardback)*. CRC Press.

Miller, E. and M. Roorda (2003). Prototype model of household activity-travel scheduling. *Transportation Research Record: Journal of the Transportation Research Board 1831*, 114–121.

Mohammadian, A., M. Javanmardi, and Y. Zhang (2010). Synthetic household travel survey data simulation. *Transportation Research Part C: Emerging Technologies 18*(6), 869–878.

Mohammadian, A. and E. Miller (2003). Empirical investigation of household vehicle type choice decisions. *Transportation Research Record: Journal of the Transportation Research Board 1854*(-1), 99–106.

Mohammadian, A. and Y. Zhang (2007). Investigating transferability of national household travel survey data. *Transportation Research Record: Journal of the Transportation Research Board* (1993), 67–79.

Moons, E. (2005). *Modeling Activity-Diary Data: Complexity or Parsimony? PhD thesis.* Ph. D. thesis, Faculteit Wetenschappen, Limburgs Universitair Centrum,Diepenbeek, Belgium.

Moriarity, C. and F. Scheuren (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics 17*(3), 407–422.

Moriarity, C. and F. Scheuren (2003). A note on rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics 21*(1), 65–73.

Nagel, K., R. Beckman, and C. Barrett (2003). Transims for transportation planning. In Y. Bar-Yam and A. Minai (Eds.), *Unifying Themes in Complex Systems: Proceedings of the Second International Conference on Complex Systems*, Volume 2, pp. 437–444. Oxford: Westview Press.

Nakamya, J., E. Moons, and W. Geert (2009). How real are synthetic populations? In *the proceedings of the 57th Session of the International Statistical Institute Conference*, Durban, South Africa.

Nakamya, J., E. Moons, S. Koelet, and G. Wets (2007). Impact of data integration on some important travel behavior indicators. *Transportation Research Record: Journal of the Transportation Research Board 1993*, 89–94.

Nakamya, J., E. Moons, P. Konstantinos, and G. Wets (2010). Creating synthetic populations for microsimulation models of activity and travel demand through different procedures: A comparative approach. *Submitted for publication.*

Nakamya, J., E. Moons, and G. Wets (2007a). Combining survey data from different studies to simulate a local travel survey sample data set: an application to the flemish region. In *the proceedings of the 11th World Conference on Transportation Research*, Berkeley, USA.

Nakamya, J., E. Moons, and G. Wets (2007b). Simulating travel duration data for flanders. In *the proceedings of the 56th Session of the International Statistical Institute, online on www.isi2007.com.pt*, Lisboa, Portugal. ISI.

Nakamya, J., E. Moons, and G. Wets (2008). Comparison between enriched travel data and the original survey data by means of a model based approach. In *the proceedings of the 8th session of the International Steering Committee on Travel Survey Conferences (ISCTSC)*, Annecy, France.

Nakamya, J., E. Moons, and G. Wets (2009). Investigating the applicability of simulated local travel data: A case study on the flemish region of belgium. *Submitted for publication after revision.*

Nakamya, J., E. Moons, and G. Wets (2010). Utilizing data integration techniques to enrich travel surveys: An evaluation of the quality of the resultant synthetic data. *Submitted for publication.*

Neter, J., M. H. Kutner, W. Wasserman, and C. J. Nachtsheim (1996). *Applied Linear Regression Models* (3 ed.). McGraw-Hill/Irwin.

Nolan, A. (2010). A dynamic analysis of household car ownership. *Transportation Research Part A: Policy and Practice 44*(6), 446–455.

Norman, P. (1999). Putting iterative proportional fitting on the researcher's desk. Working Paper 99/03, United Kingdom.

Nuyts, E. and E. Zwerts (2001). Onderzoek verplaatsingsgedrag stadsgewest hasselt- genk. deel 1: Methodologische analyse. Technical report, Onderzoekscel Architectuur en Mobiliteit, Provinciale Hogeschool Limburg, Departement Architectuur, Diepenbeek.

OBrien, S. (1999). The role of data fusion in actionable media targeting in the 1990s. *Marketing and Research Today* (19), 15–22.

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models* (1 ed.). Wiley-Interscience.

Pendyala, R., R. Kitamura, A. Kikuchi, T. Yamamoto, and S. Fujii (2005). Florida activity mobility simulator: Overview and preliminary validation results. *Transportation Research Record: Journal of the Transportation Research Board 1921*, 123–130.

Pendyala, R., A. Verma, K. Konduri, and B. Sana (2009). Socio-economic and transport trends in india and the united states: a preliminary comparative study. *Transportation Letters: The International Journal of Transportation Research 1*(2), 121–146.

Petermans, A., E. Nuyts, E. Zwerts, and E. Moons (2005). Vergelijkingonderzoek verplaatsingsgedrag vlaanderen, tijdsbudgetenqute vlaanderen, onderzoek verplaatsingsgedrag belgi, socio-economische enquete& methodologie neu kontiv design.

Piela, P. and S. Laaksonen (2001). Automatic interaction detection for imputationTests with the WAID software package. In *Proceedings of Federal Committee on Statistical Methodology 2001 Conference.*

Planbureau, F. (2009). http://www.plan.be/index.php?lang=nl&TM=30&IS=61.

Pointer, G., P. Stopher, and P. Bullock (2004). Monte carlo simulation of household travel survey data for sydney, australia - bayesian updating using different local sample sizes. *Data and Information Technology* (1870), 102–108.

Potoglou, D. (2008). Transport and environment : Vehicle-type choice and neighbourhood characteristics: An empirical study of hamilton, canada. *Transportation Research Part D 13*(3), 177–186.

Potoglou, D. and P. S. Kanaroglou (2008). Modelling car ownership in urban areas: a case study of hamilton, canada. *Journal of Transport Geography 16*(1), 42–54.

Pritchard, D. R. and E. J. Miller (2009). Advances in agent population synthesis and application in an integrated land use / transportation model. In *TRB 88th Annual Meeting Compendium of Papers DVD*, Washington, D.C.

Rassler, S. (2002). *Statistical Matching A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches.* New York: Springer.

Reiter, J. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics 21*(3), 441–462.

Rilett, L. (2001). Transportation planning and transims microsimulation model: Preparing for the transition. *Transportation Research Record: Journal of the Transportation Research Board 1777*, 84–92.

Roberts, A. (1994). Media exposure and consumer purchasing: an improved data fusion technique. *Marketing and Research Today* (22), 159–172.

Rodgers, W. L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics 2*(1), 91–102.

Rohaly, J., A. Carasso, and M. A. Saleem (2005). The Urban-Brookings tax policy center microsimulation model: Documentation and methodology for version 0304. Technical report, Tax Policy Center, Washington, DC.

Romilly, P., H. Song, and S. Liu (2001). Car ownership and use in britain: a comparison of the empirical results of alternative cointegration estimation methods and forecasts. *Applied Economics 33*(14), 1803–1818.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rubin, D. and N. Thomas (1992). Characterizing the effect of matching using linear propensity score methods with normal-distributions. *Biometrika 79*(4), 797–809.

Rubin, D. and N. Thomas (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics 52*(1), 249–264.

Ruggles, N. and R. Ruggles (1974). Strategy for merging and matching microdata sets. *Annals of economic and social measurement 3*(2), 353–371.

Ruggles, N., R. Ruggles, and E. Wolff (1977). Merging microdata - rationale, practice and testing. *Annals of economic and social measurement 6*(4), 407–428.

SAS Institute Inc. (2004). *SAS/ETS 9.1 Users Guide.* SAS Institute, Inc., Cary, N.C.

SAS Institute Inc. (2008). *SAS/STAT 9.2 Users Guide.* SAS Institute, Inc., Cary, N.C.

Shumway, R. H. and D. S. Stoffer (2005). *Time Series Analysis and Its Applications*. Springer.

Staff, U. T. I. (2008). Census bureau home page. http://www.census.gov/. The Census Bureau Web Site provides on-line access to our data, publications, and products.

Steg, L., C. Vlek, and G. Slotegraaf (2001). Instrumental-reasoned and symbolic-affective motives for using a motor car. *Transportation Research Part F: Traffic Psychology and Behaviour 4*(3), 151–169.

Stoffer, D. S. and C. M. C. Toloi (1992). A note on the LjungBoxPierce portmanteau statistic with missing data. *Statistics & Probability Letters 13*(5), 391–396.

Stopher, P., P. Bullock, J. M. Rose, and G. Pointer (2003). Simulating household travel survey data in australia: Adelaide case study. *Road and Transport Research 12*(3), 29–44.

Stopher, P., C. FitzGerald, and M. Xu (2007). Assessing the accuracy of the sydney household travel survey with GPS. *Transportation 34*(6), 723–741.

Stopher, P., S. Greaves, and M. Xu (2005). Using national data to simulate metropolitan area household travel data. *Journal of Transportation and Statistics 8*(3).

Stopher, P. and P. Jones (2003). Summary and future directions. *Transport Survey Quality and Innovation*, 635–646.

Studiedienst Vlaamse Regering (2009). Studiedienst van de vlaamse regering. http://aps.vlaanderen.be/.

Sundararajan, A. and K. Goulias (2003). Demographic microsimulation with demos 2000: Design, validation, and forecasting. In K. Goulias (Ed.), *Transportation Systems Planning: Methods and Applications*, Chapter 14, pp. 1–23. Boca Raton, FL: CRC Press.

Szalai, A. (1972). The uses of time: Daily activities of urban and suburban populations in twelve countries. Technical report, The Hague, Mouton.

Tanner, J. (1962). Forecasts of future numbers of vehicle in great britain. *Roads and Construction XL*, 263–274.

Tanner, J. C. (1977). Car ownership trends and forecasts. Transport and road research laboratory report LR 799, Transport and Road Research Laboratory,Crowthorne.

Timmermans, H., T. Arentze, and C. Joh (2002). Analysing space-time behaviour: new approaches to old problems. *Progress in Human Geography 26*(2), 175–190.

Ton, T. and D. Hensher (2003). Synthesising population data: The specification and generation of synthetic households in TRESIS. In *the 9th World Conference of Transport Research*, UK. Elsevier, Oxford.

Train, K. (1980). A structured logit model of auto ownership and mode choice. *Review of Economic Studies 47*(2), 357–370.

Train, K. and M. Lohrer (1983). Vehicle ownership and usage: an integrated system of disaggregate demand models. Washington DC.

Van Der Puttan, P., J. N. Kok, and A. Gupta (2002). Data fusion through statistical matching. Technical Report Paper 185, MIT Sloan School of Management, MIT Sloan School of Management.

Veldhuisen, J., H. Timmermans, and L. Kapoen (2000, Jan). Microsimulation model of activity-travel patterns and traffic flows: Specification, validation tests, and monte carlo error. *Transportation Research Record: Journal of the Transportation Research Board 1706*, 126–135.

Vovsha, P., E. Petersen, and R. Donnelly (2002). Microsimulation in travel demand modeling: Lessons learned from the new york best practice model. *Transportation Research Record: Journal of the Transportation Research Board 1805*, 68–77.

Vovsha, P., E. Petersen, and R. Donnelly (2003). Explicit modeling of joint travel by household members: Statistical evidence and applied approach. *Transportation Research Record: Journal of the Transportation Research Board 1831*, 1–10.

Wagner, J. (2001). The causal effects of exports on firm size and labor productivity: First evidence from a matching approach. Hamburgisches Welt-Wirtschafts-Archiv Discussion Paper 155, Hamburg Institute of International Economics, Hamburg, Germany.

Wets, G. (2005). An Activity–Based approach for surveying and modeling travel behavior. Project proposal, strategic basic research funded by institute for the promotion of innovation by science and technology in flanders, Transportation Research Institute, Diepenbeek, Belgium.

Wets, G., K. Vanhoof, T. Arentze, and H. Timmermans (2000). Identifying decision structures underlying activity patterns: An exploration of data mining algorithms. *Transportation Research Record: Journal of the Transportation Research Board 1718*(-1), 1–9.

Whelan, G. (2007). Modelling car ownership in great britain. *Transportation Research Part A: Policy and Practice 41*(3), 205–219.

Williams, B. (2001). Multivariate vehicular traffic flow prediction: Evaluation of ARIMAX modeling. *Transportation Research Record: Journal of the Transportation Research Board 1776*(-1), 194–200.

Williams, B. and L. Hoel (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering 129*(6), 664–672.

Williamson, P., M. Birkin, and P. H. Rees (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A 30*(5), 785–816.

Williamson, P. and G. P. Clarke (1996). Estimating small-area demands for water with the use of microsimulation. In C. G (Ed.), *Microsimulation for urban and regional policy analysis*, pp. 117–148. London: Pion Ltd.

Wilson, A. and C. Pownall (1976). A new representation of the urban system for modelling and for the study of micro-level interdependence. *Area 8*, 246–254.

Winkler, W. (1995). Matching and record linkage. *Business survey methods*, 355–384.

Wolff, E. (2000). Recent trends in wealth ownership, 19831998. Working Paper 300, The Levy Economics Institute of Bard College.

Wong, D. W. S. (1992). The reliability of using the iterative proportional fitting procedure. *Professional Geographer 44*(3), 340–348.

Yaffee, R. A. (2000). *An Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS* (1 ed.). Academic Press.

Yagi, S. and A. Mohammadian (2008). Modeling daily activity-travel tour patterns incorporating activity scheduling decision rules. *Transportation Research Record: Journal of the Transportation Research Board 2076*, 123–131.

Ye, X., K. C. Konduri, R. M. Pendyala, B. Sana, and P. Waddell (2009). Methodology to match distributions of both household and person attributes in generation of synthetic populations. In *TRB 88th Annual Meeting Compendium of Papers*.

Zhang, Y. and A. Mohammadian (2008a). Bayesian updating of transferred household travel data. *Transportation Research Record: Journal of the Transportation Research Board* (2049), 111–118.

Zhang, Y. and A. Mohammadian (2008b). Microsimulation of household travel survey data. In *Proceedings of the 87th annual meeting of the Transportation Research Board*, Washington, D.C.

Zhao, Y. and K. Kockelman (2000). Household vehicle ownership by vehicle type: application of a multivariate negative binomial model. Washington, D.C.

Zwerts, E. and E. Nuyts (2004). Onderzoek verplaatsingsgedrag vlaanderen 2. Technical Report D/2004/3241/016, Provinciale Hogeschool Limburg, Diepenbeek.

Nederlandse Samenvatting

Het creëren van synthetische data voor microsimulatiemodellen is een relatief nieuw onderzoeksdomein. Ondanks de momentele hoge vraag naar synthetische data in het domein van vervoer, is er weinig gekend over de methodes om deze data te verzamelen. Dit onderwerp heeft zeer recent heel wat aandacht gekregen. Zo werden er recent meerdere benaderingen voorgesteld om de kloof tussen de beschikbare data en de data die werkelijk nodig is om verschillende onderzoeksdoelen te bereiken te overbruggen. Over het algemeen vertrekken de voorgestelde benaderingen van de verschillende vormen van beschikbare data om op basis hiervan synthetische data te creëren.

De meeste nieuwe methodes die voorgesteld werden voor populatiesynthese worden nog niet begrepen door de onderzoeksgemeenschap aangezien zij nog in een vroeg stadium verkeren. Bijgevolg aarzelen beleidsmensen nog om deze methodes toe te passen. De keuze van de te gebruiken methodiek wordt gewoonlijk gedicteerd door verschillende pragmatische overwegingen, zoals bijvoorbeeld de specifieke wetenschappelijke doelstellingen, rekenkundige complexiteit, alsook de omvang en de aard van de data die gegenereerd moet worden. Omdat deze keuze de onderzoeker - om nog te zwijgen van de vakman - afschrikt, wil het onderzoek naar de verschillende procedures die hierin overwogen worden extra nuttige informatie verschaffen betreffende hun gebruik. Deze thesis beoogt een beter begrip van populatiesynthesemethodes door de toepassing van de methodes op verschillende onderzoekproblemen waarvoor nieuwe data gecreëerd of gebruikt moet worden. Een groot deel van deze thesis is daarom gewijd aan het evalueren en beoordelen van verschillende methodes bij verschillende scenario's. Het onderzoek wil bestaande werkwijzen tonen, evalueren en wijzigingen voorstellen, met als doel nuttige en belangrijke informatie betreffende hun toepassing te verstrekken. Naast het benadrukken van de relatieve verdiensten van de diverse technieken waarover de onderzoeker beschikt, lost de thesis ook praktische problemen op, inclusief de creatie van synthetische populaties voor Vlaanderen. Dit wordt bereikt door een gentegreerd model dat hier wordt voorgesteld om synthetische populaties voor Vlaanderen te genereren. De data die door het model wordt gegenereerd, wordt momenteel gebruikt voor microsimulatie in een operationeel *FEATHERS* model (Forecasting of Evolutionary Activity-Travel of Households and their Environmental RepercussionS). *FEATHERS* is essentieel als leidraad voor en ter ondersteuning van het vervoersbeleid in Vlaanderen en België.

De thesis bestaat uit acht hoofdstukken. Hierin worden verscheidene onderwerpen behandeld, zoals: data-integratie, het simuleren van gedesaggregeerde synthetische tripdata op individueel niveau, modellen die enkele tripgerelateerde variabelen voorspellen en het creëren van synthetische data op microniveau voor de populatie van Vlaanderen in toekomstige jaren.

Hoofdstuk 1 bevat een inleiding tot dit onderzoek, gevolgd door Hoofdstuk 2, waarin een gedetailleerde evaluatie gegeven wordt van de datareeksen die in deze thesis worden gebruikt.

Het eerste probleem dat in deze thesis behandeld moet worden is data-integratie; het onderwerp van Hoofdstuk 3. Statistische matching vertegenwoordigt een rijke verscheidenheid aan toepassingen in domeinen zoals microsimulaties, marketing en officiële statistieken. Statistische matching wordt hier toegepast om de beschikbare tripdata van Vlaamse huishoudens aan te vullen met informatie uit een tijdsbudgetonderzoek. Deze techniek stelt ons in staat om nieuwe gegevens toe te voegen aan bestaande onderzoeksgegevens. De matching komt tot stand door middel van enkele gemeenschappelijke variabelen tussen een huishoudelijke verplaatsingsdataset en onderzoeksgegevens naar tijdsgebruik. Wanneer men het resulterende synthetische bestand vergelijkt met de oorspronkelijke data, wordt het duidelijk dat statistische matching resultaten kan opleveren die wezenlijk vergelijkbaar zijn met werkelijke data. Basisstatistieken zoals het gemiddelde en de standaardafwijking worden vrij goed behouden. Op basis van deze bevindingen lijkt het veelbelovend om de statistische gelijkschakeling te gebruiken als integratie-instrument, om tripgegevens aan te vullen. Deze praktische test wijst ook enkele moeilijkheden aan met betrekking tot gegevensvoorbereiding. Hiervoor worden richtlijnen voorgesteld. Gezien de relatie tussen variabelen hier slechts onderzocht wordt op basis van n enkel experiment, moet de van de resultaten nog onderworpen worden aan verder onderzoek. Toekomstig onderzoek kan ook gevoerd worden naar de integratie van informatie van meerdere bestanden door middel van statistische matching waarbij externe informatie gebruikt wordt die momenteel niet beschikbaar is. Dit zal leiden naar een betere overeenstemming en validatie van de veronderstelling van voorwaardelijke onafhankelijkheid.

Een andere uitdaging die in deze thesis wordt aangegaan is het scenario om synthetische, gedesaggregeerde tripgegevens op individueel niveau te simuleren. Dit komt aan bod in Hoofdstuk 4. Een simulatiebenadering gebaseerd op Stopher et al. (2003) wordt gebruikt voor deze taak. Deze benadering wordt verder ontwikkeld door de voorstelling van een alternatieve groeperingstechniek en een uitbreiding van de methode zodat deze een diepgaande validatiebenadering bevat. Deze nieuwe aanpak wordt getest op de case studie Vlaanderen en de resultaten worden vergeleken met de resultaten, verkregen uit de vorige aanpak. In de applicatie ligt de nadruk op de simulatie van tripfrequenties en duurtijden. Een overtuigende case studie beklemtoont de nauwkeurigheid van simulatiemethodes. Er worden resultaten voorgesteld van de modelleringsaanpak waar het Poisson regressiemodel gefit wordt aan zowel werkelijke als gesimuleerde data. Het is erg interessant om op te merken dat de parameterschattingen die verkregen werden door gesimuleerde data over het algemeen niet afwijken van de schattingen die van werkelijke data afkomstig zijn. Afgezien van enkele waargenomen verschillen, zijn de parameterschattingen over het algemeen consistent in richting en omvang met gelijkaardige schattingen van standaardfouten. Onderzoek toont aan dat het bereik van de betrouwbaarheidsintervallen voor de respectievelijke parameters ook vrij gelijkaardig zijn. De resultaten van deze applicatie leiden tot een interessante vaststelling, nl. dat gesimuleerde reisdata werkelijke reisdata kunnen nabootsen.

Verschillende benaderingen om gebruik te kunnen maken van bepaalde belangrijke gegevens in deze studie, worden voorgesteld in Hoofdstuk 5. Tijdreeksmodellen worden gebruikt om het wagenbezit in België tot het jaar 2012 te voorspellen. Terwijl een groot aantal onderzoeksinitiatieven betreffende autobezit gebruik maken van discrete keuzemodellen, wijkt het gebruik van de data in deze studie nogal af, waardoor het noodzakelijk is om andere modellen voor te stellen. Over het algemeen hebben de modellen die gebruikt worden voor het huidige probleem tot doel tendensen te isoleren van onregelmatige variatie en de gevonden patronen in de voorspelling op te nemen. Het Holt's lineaire exponentiële smoothing model en de Autoregressieve Moving Average (*ARMA*)-aanpak proberen de huidige en toekomstige waarden van de afhankelijke evariabele te verklaren als gewogen gemiddelde van de vroegere waarden van elke variabele. Er wordt ook onderzoek verricht naar een Box-Tiao model dat autocorrelatie corrigeert door het beschrijven van foutmeldingen van het lineaire regressiemodel door middel van een *ARMA* proces. De Box-Tiao methode neemt niet alleen de vorige waarden en residuen van de reeks in rekening, maar biedt ook de mogelijkheid om de afhankelijke variabele te modelleren, gebruiken makend van de huidige en vroegere waardes van andere reeksen die de responsvariatie kunnen verklaren. Uiteindelijk wordt het Box-Tiao model voorgesteld om het autobezit te voorspellen. Analyses betreffende validatie en modelvergelijking tonen aan dat dit model superieur is t.o.v. andere modellen, en nauwkeurigere voorspellingen van de voorspellingsperiode oplevert. Over het algemeen geven de resultaten aan dat de vraag naar autobezit in België zal blijven stijgen tijdens het volgende decennium. De resultaten van dit onderzoek zijn belangrijk voor het openbaar beleid. Daarom worden transportplanners in België in hun zoektocht naar duurzaam vervoer geconfronteerd met de uitdaging tegenmaatregelen te bedenken die autobezit en gebruik van waarschijnlijk vervuilende auto's inperken. De impact van een toenemende motorisatie en infrastructuuruitbreiding op het milieu, de gezondheid, de energiebehoefte en gebruik van ruimte vereist bijgevolg aandacht. Een andere interessante kwestie is dat men niet volledig kan uitsluiten dat de vraag naar auto's verzadigd kan geraken. Het model dat hier voorgesteld wordt toont geen dergelijke patronen aan tijdens de voorspellingsperiode. In de toekomst kan het interessant zijn om keuzedrag niet enkel te modelleren volgens bezit en gebruik, maar ook volgens wagentype, aangezien de kilometerstand en het emissiecoëfficiënt van wagens aanzienlijk kan verschillen naargelang motorcapaciteit en leeftijd. Het zou interessant kunnen zijn om na te gaan of energieverbruik en emissies wezenlijk variëren tussen de verschillende motorcapaciteiten en leeftijden.

De noodzaak om een autobestand op microniveau te creëren voor gebruik in *FEATHERS* leidt ook tot het voorstellen van een car mileage model in Hoofdstuk 5. Het model volgt een simulatiegebaseerde aanpak waarbij de gewenste variabelen kunnen gesimuleerd worden. Het gecreëerde autobestand bevat variabelen zoals kilometerstand, brandstoftype en leeftijd van de auto. In de toekomst kan het interessant zijn om andere variabelen zoals wagentype toe te voegen in dit bestand, wanneer de beschikbare data hun verwezenlijking toelaten. Deze thesis bevat verdere onderzoeksdoelstellingen op microniveau waarvoor oplossingen worden voorgesteld. Er is een volkstellingbestand beschikbaar voor België maar hier ontbreken enkele belangrijke variabelen die nodig zijn voor *FEATHERS*. Daarom wordt er in Hoofdstuk 5 een proportioneel kansenmodel voorgesteld om het persoonlijk inkomen te voorspellen, alsook een Logistisch regressiemodel om het rijbewijsbezit te voorspellen. Deze modellen zijn erg geschikt voor deze problemen, gezien de aard van de variabelen en zij presteerden redelijk goed.

Hoofdstuk 6 concentreert zich op de uitdaging een kader te bepalen voor het creëren van een synthetische populatie. Er werd gefocust op drie methodes die onderzocht werden met het uiteindelijke doel om synthetische populaties te genereren. Deze methodes zijn: de methode voorgesteld door Beckman et al. (1996) , de tweede methode van Guo and Bhat (2007) en het *IPU* algoritme dat voorgesteld werd door Ye et al. (2009). Aangezien populatiesynthese een nieuw onderzoeksgebied is, zijn de meeste methodes gebaat met een kritische en evaluatieve houding. De methodes worden conceptueel geëalueerd om meer duidelijkheid te verstrekken en ontbrekende formele wiskundige formuleringen worden ook voorgesteld. De ideeën die op deze manier in de thesis worden ontwikkeld worden voorgesteld om meer inzicht te verschaffen in de huidig bestaande methodologie voor het genereren van synthetische data. Een geïntegreerd model wordt ontwikkeld voor het genereren van een synthetische bevolking voor Vlaanderen. Het model bevat drie hoofdcomponenten: de *input*, *synthese* en de *output*. De *IPU* methode is geïntegreerd als deel van het *synthesecomponent*, samen met de *IPF* methode. Het *inputcomponent* werd ontworpen voor diepgaande datamanipulaties en -voorbereiding. Dit component bevat eveneens model fitting. In het *outputcomponent*, tenslotte, wordt de volledige synthetische populatie van Vlaanderen gegenereerd.

De creatie van synthetische populaties is duidelijk wanneer eenheden gedefinieerd worden op n enkel niveau, hetzij dat van huishoudens of dat van personen. Wanneer een populatie gesynthetiseerd moet worden, zowel op het niveau van huishoudens, als dat van personen, dan is het probleem minder duidelijk. Hoofdstuk 7 behandelt het probleem van de creatie van synthetische data voor de bevolking van Vlaanderen op microniveau voor de jaren 2001, 2007 en 2021. In dit Hoofdstuk, wordt er dieper ingegaan op dit topic, aan de hand van een toepassing. De methodes die in Hoofdstuk 6 aan bod kwamen worden hier uitgevoerd en toegepast, en de resultaten worden vergeleken om de sterktes en/of zwaktes van de methodologieën en de resulterende kwaliteit van de synthetische data te benadrukken. De/het *IPU* heeft minder berekeningstijd nodig in vergelijking tot Guo and Bhat (2007) , hetgeen een troef is van de eerste methode met betrekking tot de praktische toepasbaarheid. Onderzoek wijst uit dat het *IPU* algoritme bijzonder nuttig is aangezien het een aanpak levert die de vaak vervelende en computerintensieve aspecten van populatiesynthese bij het genereren van synthetische data uitsluit. Het onderzoek benadrukt dat met sommige nieuwe methodes *goede* synthetische populaties gecreëerd kunnen worden. Men stelt verder vast dat de controle van attributen op zowel huishoudelijk als persoonlijk niveau. Bovendien worden de distributies van ongecontroleerde attributen ook redelijk goed bewaard. Er is ook een stabiliteit in de resultaten van verschillende synthetische populaties die aan de hand van deze methodes werden gegenereerd. Het resultaat van de vergelijking van de distributies van zowel ongecontroleerde huishoudens en attributen op persoonlijk niveau met waargenomen tegenhangers vormt een belangrijke bijdrage tot het onderzoeksgebied van de creatie van synthetische populaties omdat in alle eerdere onderzoeken enkel gecontroleerde attributen met elkaar vergeleken werden. De bevindingen sporen aan tot verder onderzoek naar hoe een goede synthetische populatie de voorspelling van de vervoersvraag verbetert. Een natuurlijke aanvulling van dit onderzoek in toekomstige exploraties zou kunnen bestaan uit de toepassing van het *FEATHERS* model op de Vlaamse synthetische populatie om zo het aandeel in de vervoersparticipatie te vergelijken met het aandeel dat gerapporteerd word in enquêtes of het aandeel dat gegenereerd wordt wanneer *FEATHERS* toegepast wordt op werkelijke data.