Combining Survey Data from Different Studies to Simulate a Local Travel Survey Sample Data Set: An Application to the Flemish Region

Juliet Nakamya
PHD Student
Hasselt University
Transportation Research Institute
Wetenschapspark 5, bus 6
B-3590 Diepenbeek
Belgium
T: +32 (0) 11 26 91 32
F: +32 (0) 11 26 91 99
E-mail: juliet.nakamya@uhasselt.be

Elke Moons
Doctor
Hasselt University
Transportation Research Institute
Wetenschapspark 5, bus 6
B-3590 Diepenbeek
Belgium
T: +32 (0) 11 26 91 26
F: +32 (0) 11 26 91 99
E-mail: elke.moons@uhasselt.be

Geert Wets
Professor
Hasselt University
Transportation Research Institute
Wetenschapspark 5, bus 6
B-3590 Diepenbeek
Belgium
T: +32 (0) 11 26 91 58
F: +32 (0) 11 26 91 99
E-mail: geert.wets@uhasselt.be

Total number of words: 5,500
Re-submission date: April 30, 2007

Submitted for presentation at the 11th World Conference on Transportation Research, June 2007

ABSTRACT

The aim of this paper is enriching the Flemish Household Travel Survey (*FHTS*) data with local socio-demographic data, available from the National Institute of Statistics (*NIS*) and further incorporating time-use data, available from the 'Time-use of the Flemish people' survey into this framework in order to make more reliable simulations of travel data. The travel attributes to be simulated are examined and households/individuals classified into groups that exhibit similar ranges of selected travel attributes. Using these groupings, distributions of the selected travel attributes are produced, which then become the basis of the simulation. Future research will mainly focus on in-depth validation of the outputs of the simulation process, investigation of the stability of results from different simulation runs and further improvements involving local data updates. It is anticipated that this approach will enable Flanders to develop a local travel data set and estimate travel-demand models at a fraction of the cost of conducting a traditional household travel survey.

Keywords: *Classification, Data integration, Simulation*

1 INTRODUCTION

Travel surveys are presently one of the most important ways of obtaining the critical information needed for transportation planning and decision making. These surveys are used to collect current information about the demographic, socio-economic, and trip-making characteristics of individuals and households as well as furthering our understanding on travel in relation to the choice, location, and scheduling of daily activities. This enables enhancement of travel forecasting methods and improves the ability to predict changes in daily travel patterns in response to existent social and economic trends as well as new investments in transportation systems and services.

The growing need for timely, quality, and large amounts of data and information from national statistical agencies has increased continuously over the years. Provision of large quality data on travel demand related to the socio-demographic and travel characteristics of individuals and households, largely relies on household travel surveys (*HTS*). However, it goes without mention that *HTS* are notoriously expensive and require an appreciable amount of time to plan and execute, despite the current state of increasingly tight budgets. While methodological and technological survey techniques become increasingly refined, high unit costs and public resistance will continue to plague future survey efforts. It is fair to believe that even when more advanced and recent technologies such as the global positioning system (Murakami and Wagner, 1999; Draijer *et al.*, 2000; Murakami *et al.*, 2000) and personal digital assistant (Murakami *et al.*, 2000; Janssens, *et al.*, 2004) are used, the final total cost will only become higher. Another, yet big problem faced in conducting high-quality travel surveys today is non-participation. Researchers are now becoming increasingly concerned about the high response burden imposed on respondents especially due to the fact that response rates are dropping dramatically. The implications of all these problems on the quality and representativeness of the resulting data are startling.

Combining data from different surveys can be a plausible option in an effort to reduce respondent burden and survey costs. A practical solution is to exploit as much as possible all the information already available in different data sources, that is, to carry out a statistical integration of information that has already been collected. While a significant amount of work has been done 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. Integration of data from different sources can be performed by means of three different methodologies: record linkage, merging and statistical matching. The record linkage and merging techniques are substantially different from the statistical matching problem. They are designed to link the same units from two or more different files. Merging requires error-free matching variables, while record linkage is a statistical decision procedure that can be used when matching-variables are affected by errors. Both techniques require that the sets of units in the two sources overlap. Statistical matching, which is also the technique that was used for the combined data used in this paper, targets providing joint information on variables observed in different sources. It faces the problem of integration when the files lack unit identifiers or do not contain the same units.

Furthermore, the problems faced in conducting travel surveys together with the extensive data needs of new disaggregate-based approaches to forecasting travel, warrant developing techniques for augmenting or replacing existent data collection techniques with artificial/synthetic data. Simulating *HTS* data (Greaves and Stopher, 2000; Stopher *et al.*, 2003; Pointer, *et al.*, 2004) is a relatively new field of research with many potential benefits. The benefits of this approach would be enormous to all actors who utilize *HTS* data. It could potentially provide a very low cost technique for generating a local sample of many additional households in comparison to collecting household travel data. Simulated data would also conceivably form a component of the *HTS* currently being undertaken, and facilitate these surveys to reduce sample size requirements without necessarily compromising the quality of the planning activities supported by these data. However, this area of research is still in its infancy and significant work still has to be done. Such work should be geared towards developing a state-of-the-art technique, testing and also establishing the clear role for household travel data simulation.

Ideally, travel behavior is essentially a result of complex interactions and correlations between households, individuals, their residential area choices and the existent transportation system. Simulation (Greaves, 2006; Stopher *et al.*, 2003; Janssens *et al.*, 2004) theoretically overcomes this problem by applying widely-used techniques to replicate observed behavior, which implicitly encapsulates these interactions. The complete nature of these interactions would otherwise be challenging to fully capture in analytical models. Simulation, further by its nature, captures the variability in behavior, which is suppressed by the traditional measures of central tendencies. Simulation also enables us to work with much larger samples. However, it is evident that any systematic biases or unusual behavior in the base data would also be reflected in the simulated data. If the existing data does not adequately capture the behavior of the representative sample, then, neither will the simulated data. In reality, *HTS* may cover only a small sample of the population thus failing to be representative of the target population. This makes combination of *HTS* data with data from other related surveys, a vital option so as to obtain larger data base samples.

The main aim of this paper is to simulate a local travel data set using household travel data enriched with time use survey data. The data available in this study include data from the Flemish Household Travel Survey (*FHTS*) carried out in Belgium in 2000 (Zwerts and Nuyts, 2004), the Flemish Time Use Survey (*FTUS*) also carried out in Flanders in Belgium in 1999 (Glorieux, 2000) and additionally, the Socio-Economic population census (*SEE*) data of 2001 conducted in Belgium (Statistics Belgium, 2001) are also available. The *FHTS* and the *FTUS* data are combined (Nakamya *et al.*, 2007) using the *SEE* as the base data. Categories of individuals that exhibit similar ranges of trip rates conditional on the travel mode, are developed using the Classification and Regression Trees method. Using these groupings, distributions of the trip rates are produced, which then become the basis of the simulation. The results of the simulation are then compared with the actual survey results.

The remainder of this paper is organized as follows. Section 2 describes the surveys that resulted in the data available in this study. In Section 3, the methodology used in this study is laid out. The results are then presented and discussed in Section 4 and finally, Section 5 gives the concluding remarks and some directions for further research are presented.

2 DATA

The survey data used in this study arise from two surveys: the Flemish Household Travel Survey (*FHTS*) carried out in 2000 (Zwerts and Nuyts, 2004) and the Flemish Time Use Survey (*FTUS*) carried out also in Flanders, Belgium in 1999 (Glorieux, 2000). Table 1 gives a comparison of the sample design of the *FHTS* and the *FTUS* surveys.

<Insert Table 1 about here>

The *FHTS*, which is the major survey of interest in this study, was carried out among the Flemish citizens. The *FHTS* field work was carried out 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 asked to fill in an individual questionnaire and also to keep a travel diary for two days. The individual questionnaire included socio-demographic variables as well as travel-related variables. In the travel diary, respondents recorded their travel activities, modes of transport, duration, location, company of others when traveling and search for car parking. Further data was collected from these households using household questionnaires. This survey had a response rate of 32% of the households. The second survey, *FTUS,* was carried out by the *Tempus Omnia Revelat* research group of the Free University of Brussels amongst the Flemish citizens. The fieldwork took place between April 15 and October 30, excluding the period between the 15th of July and the 1st 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 diary for a full 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. Further more, respondents were asked their opinion about different social issues. A response rate of 28% of the individuals was obtained in this survey.

3 METHODOLOGY

3.1 Data Integration

Data integration effected through statistical matching is initiated by two or more samples, one usually larger than the other with a negligible overlap of units (*e.g.* individuals) in both samples. D'Orazio *et al.*, (2006) tackled the statistical matching problem providing a consistent maximum likelihood estimator of the elements characterizing uncertainty. There are two broad groups of objectives for statistical matching: the *micro* and *macro* objectives. The *micro* approach is obtained when interest is essentially in integrating the database at unit level, and the *macro* approach, when most interest is in the aggregates. Statistical matching methodologies should be chosen according to these two previous objectives.

The Iterative Proportion Fitting (*IPF*) method (Norman, 1999) is a well established technique with the theoretical and practical considerations behind the method thoroughly explored and reported. The method was developed for combining information from two or more sets of data (Bishop *et al.*, 1995). It uses the population or the larger sample margins to update the information at cell frequency level. However, in this study, since socio-demographic population data is fully available, the internal population frequency cell values for the respective classes of interest are directly used to calculate the weights instead of using population marginal values.

In dealing with data integration, a great problem encountered, is that of harmonizing the different data sources. This may tend to be somewhat expensive and time consuming. This is especially the case when the surveys use different scales of measurement for common variables and different approaches to deal with different survey aspects. The two survey data sets (*FHTS* and *FTUS*) available here, are separately cleaned and adjusted to make them compatible with each other. Since the two samples were selected from the same population, they were each weighted with respect to the Flemish population data (*SEE*) to ensure representativity (Nakamya *et al.* 2007). The two sets of survey data were further combined on some socio-demographic characteristics and some common travel characteristics. Comparison of the combined data set with the original *FHTS* data set on some travel behavior indicators was then carried out. In this current study, we utilize the *FHTS* data and these combined data and focus on only the respondents who participated in travel activities.

3.2 Development of the Simulation Procedure

The aim of this phase is to simulate a travel survey data set for a target sample. Theoretically, one should be able to reproduce (within an acceptable error range) the collected data and build models that are similar to a real survey. Preparation for this procedure, involves categorizing individuals/households and developing distributions from which samples are drawn. Using homogeneous groups of individuals rather than households has merit, particularly with respect to mode choice (Supernak et al., 1983). While some researchers (Greaves, 2000) have chosen to work at the household level in conducting related research, others have chosen to simulate data for individuals (Axhausen and Herz, 1989; Kulkarni and McNally, 2001; Raney and Nagel, 2003). An argument is put forward that, if the goal of the simulation is to provide a household travel survey dataset, which is realistic and plausible, this has (ultimately) to be done at the level of individual households members (Greaves, 2006). In this study, we choose to work at the individual level in simulating the number of trips by mode of travel used.

3.2.1 Categorizing Individuals

An initial step geared towards setting up the simulations is to categorize individuals into relatively homogeneous groupings with respect to the number of trips made by mode of travel. The data that we use here to formulate these categories, as explained earlier, are obtained after combining the *FHTS* and the *FTUS* data with the Belgian socio-demographic census data as the base data. The Classification and Regression Tree (CART) method, a computationally intensive exploratory classification tool proposed by Breiman *et al.*, (1984) is used to this effect. This method involves three major stages:

1. Grow an overly large tree to capture all potentially important splits;
2. Prune the tree back to the root node to create a hierarchy of sub trees;
3. Select an optimal-sized tree from this sequence using an independent holdout sample or cross-validation. As suggested by Breiman *et al.*, (1984) “optimal” 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.

The CART method, a nonparametric and nonlinear technique, involves binary recursive partitioning of the data with respect to the dependent variable of interest. The algorithm works in the forward direction implying that once a node is split, it cannot change. At each node, all predictor variables are evaluated to determine the best groupings based on their reduction in the residual sum of squares (improvement score). The independent variable with the largest improvement score is selected for the split, and then the process continues until some user-defined stopping criteria are met. To explain the technique further, a tree consists of different layers of nodes. It 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 become in turn parent nodes. This recursive partitioning algorithm continues until a node is terminal and has no offspring (determined by a stopping criterion). Nodes in deeper layers become more and more homogeneous, less 'impure', with respect to the response variable. 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. The 'goodness' of a split can be defined as the reduction in impurity

$$\Delta I(\tau) = i(\tau) - P(\tau_L)i(\tau_L) - P(\tau_R)i(\tau_R)$$

with $i(\tau)$ denoting the impurity of the node τ and $P(\tau_L)$ the probability that a subject falls into the left daughter node τ_L of node τ. Likewise, $P(\tau_R)$ defines the probability corresponding to the right node. A popular example of such an impurity measure is the entropy measure

$$i(\tau) = -p_\tau \log(p_\tau) - (1 - p_\tau)\log(1 - p_\tau)$$

with $p_\tau = P(Y = 1/\tau)$, where *Y* is the response. In the pruning process, the initial tree is then pruned recursively, leading to a sequence of pruned and nested sub trees. From this sequence of trees, we choose the sub tree with *g* terminal nodes. The procedure of cross validation can be used in making a choice of the right size tree. It is based on optimal proportion between the complexity of the tree and misclassification error. This task is achieved through the cost-complexity function

$$R_\alpha(T) = R(T) + \alpha(\tilde{T}) \rightarrow \min_T$$

where $R_\alpha(T)$ is the misclassification error of the tree *T*; $\alpha(\tilde{T})$ the complexity measure which depends on $\tilde{T}$, the total sum of terminal nodes in the tree.

Travel-demand models assume that there are relationships between socio-demographic characteristics and travel characteristics, by using variables describing households / individuals, particularly, trip generation and mode-choice modeling. It is well established that total trip generation is associated with the demographic and socio-economic attributes of the traveller (Ortúzar and Willumsen, 2006). In the study at hand, the dependent variable used is the total number of trips made per day by individual respondents. The independent variables 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 mode of travel used by the respondents, "homogeneous" categories of the trip rates of individuals are developed using the regression trees methodology.

3.2.2 Developing Distributions

The next step towards setting up the simulation is developing frequency distributions from which, sampling is made for the travel characteristic of interest. Within each of the established "homogeneous" categories developed in the preceding stage, the trip rates exhibit some variation. To capture this variation therefore, discrete frequency distributions of values of trip rates were developed by recording their magnitude for each occurrence of the category in the combined data set. The frequency distributions are then re-constructed as cumulative frequency distributions with each discrete value of the attribute now falling within a particular probability range. This then provides the basis for the random sampling process used in the data simulation procedure. This procedure of developing distributions is then repeated for each category to create a "family" of cumulative frequency distributions. After this stage, the next task was to sample from these distributions so as to simulate trip rates for the target sample.

4 RESULTS AND DISCUSSION

The *FHTS* and the *FTUS* data comprise of 6,401 and 1,527 respondents respectively within the ages of 16 and 75 years. Of these, 5,148 and 1,164 respondents conducted travel activities in the *FHTS* and the *FTUS* data respectively. Consequently, the combined data set contains 6,312 travelers. The *FHTS* and the *FTUS* trip-level data files comprise of 18,125 and 4,181 trips respectively. Thus, the combined trip-level data file contains 22,306 trips.

It would be desirable to develop one categorization scheme to predict all travel attributes (purpose of travel, departure time, mode of travel and trip length). However, in this research, we develop different schemes for each dependent variable. The results reported here focus on the mode of travel trips. Some categories of the attribute *mode of travel* are re-combined to form three groups: foot/bicycle, car and other mode. The later category includes: use of public transport, which forms about 3.8% of the total trips; use of motorbikes forming about 1.8% of the total trips and lastly; the use of other modes of travel including modes that were undefined by the users. Table 2 shows the percentage number of trips conducted by individual respondents following the used mode of travel. The percentage number of trips is shown both for the *FHTS* and the combined data. These percentages are approximately equal for the *FHTS* data and the combined data across the different travel modes. Car mode forms the highest share of trip rates of 64.5%, followed by use of foot/bicycle, which counts about 23% of the total trips made.

< Insert Table 2 about here>

CART runs are completed for the three travel modes: foot/bicycle, car and other mode. In the data combination analyses, the Belgian socio-economic census data is used as the base for computing weights for the *FHTS* and the *FTUS* data. In Nakamya *et al.* (2007), the combined data obtained by integrating these *FHTS* with the *FTUS* data, are found to be more representative of the population than the original travel (*FHTS*) data. Here, the classification variables available both in the *FHTS* and the obtained combined data include: gender, age group, marital status and education level. The age range of '16-75' years was considered since the *FTUS* data comprise of only respondents between the ages of 16 and 75 years. Thus, the combined data set contains individuals within the same age range.

Table 3 shows the average number of trips made per person per day, following the socio-demographic factors in the *FHTS* and the combined data. In this study, the average number of trips is calculated per person traveling or participants in travel. Thus, the trip rates are expected to be higher than in studies (Nakamya *et al.*, 2007; Zwerts and Nuyts, 2004) with general interest in trip rates for all persons in the sample. Overall, the trip rates do not significantly differ between the *FHTS* and the combined data across all demographic characteristic groups. Respondents with college or university degree stand out as the most mobile group and those with primary school certificates are the least mobile. The overall trip rate for a traveller is 3.5% for both data sets. Regarding age, people between 35 and 54 years tend to conduct more trips than those from other age groups, but these trip rates highly reduce when people become older (55-75 years). It can also be noted (with respect to marital status) that while the trip rates for the married persons are high; the rates for the widowed group are considerably low.

< Insert Table 3 about here>

Figure 1 shows the output of a CART run obtained for car mode trips using the combined data set. The cell means, the standard deviations together with the cell sizes are displayed for the terminal nodes. This final regression tree contains eleven terminal nodes thus forming eleven 'homogeneous' categories of the segmentation results. All the four classification variables are used in the actual construction of the tree. The education level and age group of individuals are the top two variables to be split upon, which indicates the high importance of these variables for car mode trips. The mean trip rates for the final categories 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.

<Insert Figure 1 about here>

For the foot/bicycle trips, the CART 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. Considering the 'other' mode trips, five final categories are formed and the most important factor is found to be the individuals' marital status.

Basing on the established categories, the next phase of the simulation procedure is to develop frequency distributions from which to sample, for the characteristic of interest. In this paper, the travel mode trips characteristic is illustrated. For the car trips attribute, individuals have been classified into eleven categories/groups from the CART procedure. The number of the car mode trips is then recorded for each occurrence of each category in the combined data set. Thus, eleven discrete frequency distributions are formed. These distributions are reconstructed into cumulative frequency distributions, which are shown in Figure 2. For a better visibility, one individual (outlier) belonging to group 8, who conducted 17 trips is excluded from this figure.

< Insert Figure 2 about here>

Figure 3 shows the probability distribution of group 1 (primary school or Junior high school persons within 55-75 years) for car mode trips superimposed onto a Poisson distribution with mean 2.406. The Figure clearly shows that these trip counts can be assumed to follow a Poisson distribution. However in this case, since our data set does not contain zero trips (only travellers are considered), we truncate the Poisson distribution and only values other than zero are selected from the distribution.

< Insert Figure 3 about here>

Random samples are taken from the corresponding distributions for each created 'homogeneous' category. For each distribution, a sample proportional to the size of the category in the combined data set is drawn. To complete the simulation procedure, this process is then repeated for each travel characteristic of interest. As a result in this case, a complete travel survey data set for each individual is generated with regards to mode trip rates. Table 4 gives a comparison of the car mode trip rates for the combined survey data with the simulated data following the earlier generated categories. Based on the categories created from the combined data, summary statistics are also obtained using the *FHTS* data and are shown in the second column of Table 4. One general observation here is that car trip rates of the travel survey data seem to be replicated relatively well by the simulated data across the categories.

< Insert Table 4 about here>

Table 5 shows a comparison of the bike/walk mode simulated trip rates with the trip rates from the actual 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 (*FHTS* and combined survey data), although slightly higher for most of the categories.

< Insert Table 5 about here>

For the 'other' mode trip rates however (Table 6), the trip rates from the simulated data tend to differ quite markedly from the actual survey data across the categories. This is highly attributable to the small cell sizes in the combined data, on which the simulation is based. It can also be further attributed to the higher variances of the trip rates. Creation of multiple datasets further tends to mitigate these problems.

< Insert Table 6 about here>

It goes without saying that each time a sample is drawn from a distribution, a different simulated data is obtained, therefore multiple synthetic datasets can be created and aggregate statistics combined in a manner akin to a 'multiple imputation' process (Rubin, 1987).

5 CONCLUSIONS AND FURTHER RESEARCH

This paper uses data combined from a household travel survey and a time use survey to simulate a local travel survey sample dataset. Combining travel data with data from other related sources provides a larger and more representative sample of the population, which gives more reliable travel information on the population. The larger sample is valuable in prediction of travel demand and offers a good base for simulating travel data.

The results presented here generally show that the concept of simulating data from travel data is a very prospective one, which deserves further attention. A simulation procedure has been set up by initially developing homogeneous categories of individuals following the trip mode attribute for the individuals. This was done by completing Classification and Regression Tree runs for trips conducted, conditional on the mode of travel. Distributions were then formulated for each of the obtained categories, from which samples were subsequently drawn to obtain a synthetic travel data set. In general, the procedure provides results that are comparable with results from an actual travel survey. The next planned phase of this work is to simulate the other required travel data attributes including trip purpose, trip length and departure times.

Future research will mainly focus on in-depth validation of the outputs of the simulation process, investigation of the stability of results from different simulation runs and further improvements involving local data updates. It is anticipated that this approach will enable Flanders and other regions or countries, to develop a local travel data set and estimate travel-demand models at a fraction of the cost of conducting a traditional household travel survey.

REFERENCES

Angrist, J. D. and Krueger, A. B. (1992) 'The Effect of Age at School on Entry on Educational Attainment: An Application of Instrumental variables with Moments from Two Samples' *Journal of the American Statistical Association*. **87 (**418**)**.

Arellano, M. and Meghir, C. **(**1992**) '**Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets' *Review of Economic Studies*, **59(**3), pp. 537-559.

Axhausen, K.W. and R. Herz (1989) 'Simulating Activity Chains: A German Approach, *ASCE Journal of Transportation Engineering'* **115** (3), pp.316-325.

Bishop, M. M., Fienberg, S. E. and Holland, P. W. (1995) *Discrete Multivariate Analysis: Theory and Practice,* The MIT, England.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*, Wadsworth International Group, Belmont, Califonia.

D'Orazio, M., Di Zio, M. and Scanu**,** M. (2006) *Statistical Matching: Theory and Practice*, John Wiley and Sons, Inc., New York.

Draijer, G., Kalfs, N. and Perdok, J. (2000) 'Global positioning system as data collection method for travel research' *Transportation Research Record*, No. 1719, pp. 147-153.

Glorieux, I., Koelet, S. and Moens, M. (2000) 'Technisch verslag bij de tijdsbudgetenquête tor'99: Veldwerk en responsanalyse. (tor2000/43)' Vrije Universiteit Brussel, Belgium (in Dutch).

Greaves, S. P. (2006) 'Simulating Household Travel Survey Data' University of Sydney.

Greaves, S. P. and Stopher, P. R. (2000) 'Creating a Simulated Household Travel/Activity Survey–Rationale and Feasibility Analysis' *Transportation Research Record*, No.1706*,* pp. 82- 91.

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

Janssens, D., Wets, G., Brijs, T. and Vanhoof, K. (2004) 'Simulating Activity Diary Data by Means of Sequential Probability Information: Development and Evaluation of an Initial Framework' Paper Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC, January 2004.

Kulkarni, A.A. and McNally, M.G. (2001), 'A Microsimulation of Daily Activity Patterns' Paper Presented at the 80th Annual Meeting of the Transportation Research Board, Washington DC, January 2001.

Lusardi, A. (1996) ‘Permanent Income, Current Income, and Consumption: Evidence from Two Panel Data Sets’ *Journal of Business and Economic Statistics*. **14** (1).

Murakami, E. and Wagner, D. P. (1999) ‘Can using global positioning system (GPS) improve trip reporting?’ *Transportation Research C*, 7(2/3), 149.165.

Murakami, E., Wagner, D. P. and Neumeister, D. M. (2000) ‘Using global positioning systems and personal digital assistants for personal travel surveys in the United States’ Transport Surveys: Raising the Standards, Transportation Research Circular, E-008, TRB, National Research Council, Washington, D.C., III-B/1-21.

Nakamya, J., Moons, E. and Wets, G. (2007) ‘The Impact of Data Integration on Some Important Travel Behavior Indicators’ Forthcoming in the *Transportation Research Record*, journal of the Transportation Research Board.

Norman, P. (1999) ‘Putting Iterative Proportional Fitting on the Researcher's desk’ Working Paper 99/03, United Kingdom.

Ortúzar, J. and Willumsen, L. G. (2006) *Modelling Transport*, John Wiley and Sons, Inc., West Sussex, England.

Pointer, G., Stopher, P. and Bullock, P. (2004) ‘Monte Carlo simulation of household travel survey data for Sydney, Australia: Bayesian updating using different local sample sizes’ *Transportation Research Record*, No. 1870, pp. 102-108.

Raney, B. and Nagel, K. (2003) ‘Truly agent-based strategy selection for transportation simulations’ Paper presented at the Transportation Research Board Annual Meeting, Paper 03-4258, Washington, D C.

Rubin, R. B. (1987) *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons, New York.

Statistics Belgium. http://www.statbel.fgov.be/ Accessed April, 2007.

Stopher, P.R., Bullock, P. and Rose, J. (2003) ‘Simulating household travel data in Australia: Adelaide case study’ *Road and Transport Research*, **12** (3), pp. 29-44.

Supernak, J., Talvitie, A. and DeJohn, A. (1983) ‘Person-Category Trip-Generation Model’ Transportation *Research Record*, No. 944, pp. 74-83.

Szalai, A. (1972) ‘The Uses of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries’ The Hague, Mouton.

Winkler, W. E. (1995) "Matching and Record Linkage," in B. G. Cox *et al*. (ed.) *Business Survey Methods*, John Wiley and Sons, Inc., New York, pp. 355-384.

Zwerts, E. and Nuyts, E. (2004) ‘Onderzoek Verplaatsingsgedrag Vlaanderen 2 (D/2004/3241/016)’ Provinciale Hogeschool Limburg, Diepenbeek (in Dutch).

LIST OF TABLES AND FIGURES

TABLES

Table 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
N persons	7626	1533
N Households	3027	Not applicable
Sampling	Stratified sample (age of head of household)	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 - Travel Questionnaire (2 days/ retrospective)	- Individual Questionnaire - Diaries (7 days/ simultaneous)

Table 2: Percentage Number of Trips Made by Respondents Following Mode of Travel

Travel modes	*FHTS* data	Combined data
Foot/ Bicycle	23.07	22.51
Car	64.53	64.54
Other/undefined	12.39	12.95

Table 3: Average Number of Trips per Person per Day by Socio-demographic Factors

Socio-demographic characteristics	*FHTS data*	Combined data
Gender		
Male (1)	3.45	3.44
Female (2)	3.50	3.59
Age group		
16-34 years (1)	3.56	3.60
35-54 years (2)	3.66	3.75
55-75 years (3)	3.02	2.96
Marital Status		
Married (1)	3.54	3.59
Divorced (2)	3.50	3.50
Widowed (3)	2.87	2.83
Un-married (4)	3.40	3.43
Education level		
Primary school (1)	2.77	2.77
Junior high school (2)	3.16	3.19
High school (3)	3.57	3.65
College or University (4)	4.08	4.04
Overall	3.47	3.51

Table 4: Comparison of the Car mode Trip Rates between the Survey Data and the Simulated Data

Group	Categorization Scheme*	Mean (Standard deviation)		
		FHTS data	Combined Data	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)

*Category codes are as defined in Table 3

Table 5: Comparison of the Bike/Walk mode Trip Rates between the Survey Data and the Simulated Data

Group	Categorization Scheme*	Mean (Standard deviation)		
		FHTS data	Combined Data	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)

*Category codes are as defined in Table 3

Table 6: Comparison of the 'Other' mode Trip Rates between the Survey Data and the Simulated Data

Group	Categorization Scheme*	Mean (Standard deviation)		
		FHTS data	Combined Data	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)

*Category codes are as defined in Table 3

FIGURES

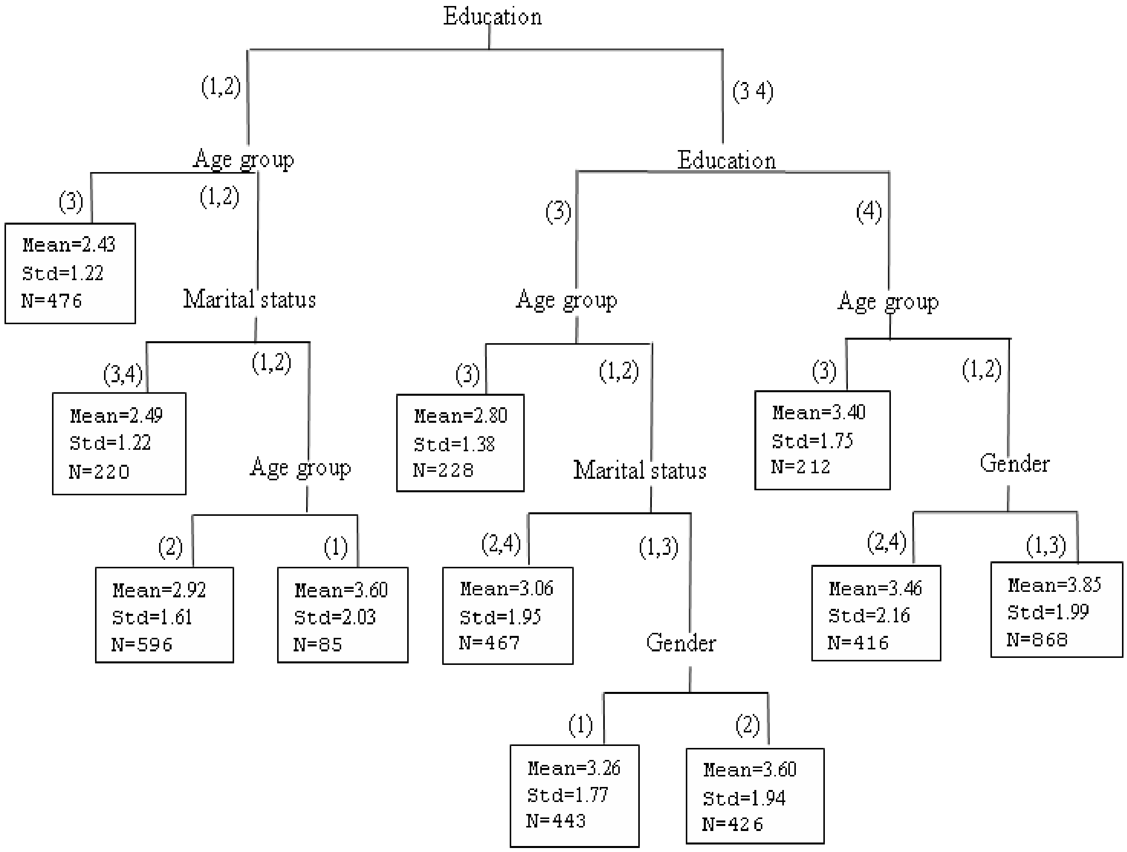

Figure 1: CART Segmentation Results for the Car mode Trips (based on the Combined Data)

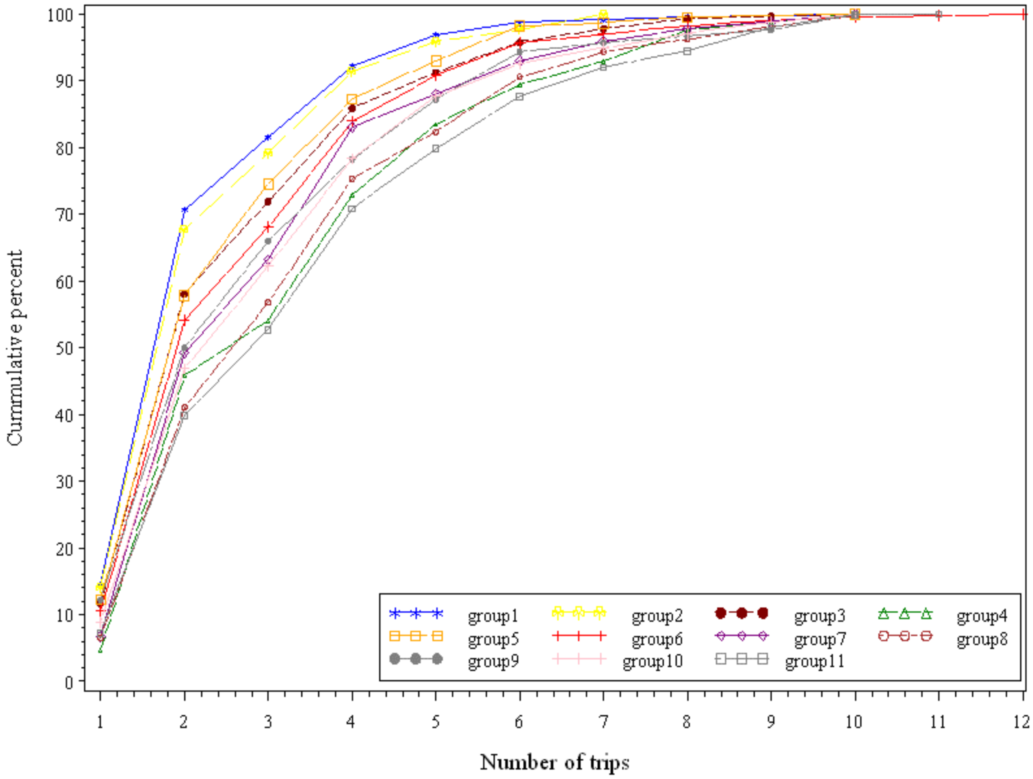

Figure 2: Cumulative Frequency Distribution of the Car mode Trips (based on the Combined Data)

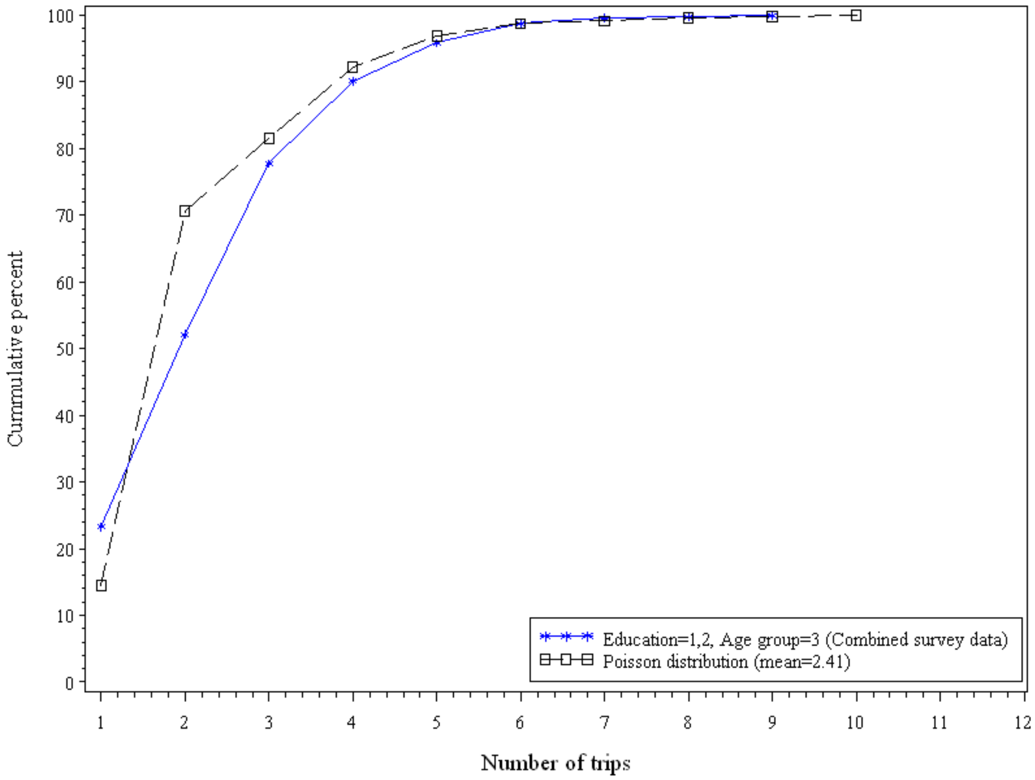

Figure 3: Cumulative Frequency Distribution of the Car mode Trips (based on the Combined Data) Superimposed onto a Poisson distribution