Simulating Travel Duration Data for Flanders

Juliet Nakamya
Hasselt University, Transportation Research Institute
Wetenschapspark 5, bus 6
B-3590 Diepenbeek, Belgium
E-mail: juliet.nakamya@uhasselt.be

Elke Moons
Hasselt University, Transportation Research Institute
Wetenschapspark 5, bus 6
B-3590 Diepenbeek, Belgium
E-mail: elke.moons@uhasselt.be

Geert Wets
Hasselt University, Transportation Research Institute
Wetenschapspark 5, bus 6
B-3590 Diepenbeek, Belgium
E-mail: geert.wets@uhasselt.be

1 Introduction

The growing need for timely, high quality and large amounts of data and information from national statistical agencies has increased continuously over the years. These data are fundamental in enhancing modern transportation planning and policy development endeavors. 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*). Nevertheless, it goes without mention that conducting these surveys is continuously characterized with numerous problems. The surveys involve high expenses, necessitate a lot of time to plan and implement and further impose a high burden on the respondents, subsequently bringing about low response rates. This often has severe consequences on the quality and representativeness of the resultant collected data. It is fair to believe that even when more advanced and recent technologies such as the global positioning system and personal digital assistant (Murakami *et al*., 2000) are used, the final total cost will only become higher. In search for plausible approaches to solve these problems, combining travel data with data from other sources becomes an attractive option. Furthermore, the success in research directed towards simulating travel data is of absolute interest in this respect. Simulating household travel survey (*HTS*) data (Greaves and Stopher, 2000; Stopher *et al.*, 2003; Pointer, *et al.*, 2004) is a relatively fresh field of research with many potential benefits. The benefits of this approach would be enormous to all actors who utilize *HTS* data and 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. In past research (Nakamya *et al.*, 2007*b*), promising results with regards to trip rates data simulation were observed. In general, the procedure provided results that were comparable with results from an actual travel survey.

The combined survey data available in this current study arise from two surveys: the Flemish Household Travel Survey (*FHTS*) carried out in 2000 and the Flemish Time Use Survey (*FTUS*), also carried out in Flanders, Belgium in 1999. The *FHTS* was carried out among 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 containing socio-demographic and travel-related variables. They also kept a travel diary for 2 days recording travel activities, modes of transport, duration, location, company of others when traveling and search for car parking on addition to filling-in household questionnaires. The *FTUS*, carried out by the Tempus Omnia Revelat research group of the Free University of Brussels targeted Flemish citizens from 16 to 75 years. Here, 1,533 people recorded all their activities in a diary for 1 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 for carrying out the activity. The *FTUS* also employed individual questionnaires recording socio-demographic variables as well as general indicators on time use and cultural participation.

This paper focuses on simulating duration travel data conditional on the purpose of travel. The combined data (Nakamya *et al.*, 2007*a*), obtained by combining the *FHTS* and *FTUS* data that both make use of the Belgian Socio-Economic population census (*SEE*) as the base data, are utilized in conducting the analyses. This combined data set was noted to offer a larger and more representative sample of the population, which gives more reliable travel information on the population, is valuable in prediction of travel demand and can also be used as a base for simulating travel data. Here, the combined data is randomly split into two sets. The bigger data (training) set is utilized to formulate categories of individuals that exhibit similar ranges of trip duration conditional on the travel purpose, using the Classification and Regression Trees method (*CART*). Distributions of trip durations are then developed using the generated groupings and these further become the basis of the simulations. The reserved smaller data (testing) set is finally used to compare results of an actual survey to the simulated travel data.

2 Methodology

In dealing with data integration (D'Orazio *et al.*, 2006), a great problem encountered, is that of harmonizing the different data sources. This exercise may tend to be somewhat expensive and time consuming. The two survey data sets (*FHTS* and *FTUS*) available here, were separately cleaned and adjusted for compatibility with each other. Since the two samples were selected from the same population, they were each weighted with respect to the *SEE* data to ensure representativity. The data were further combined on some socio-demographic characteristics and some common travel characteristics (Nakamya *et al.*, 2007*a*). In this current study, after further adjustments on the data, inserting very short trips lasting 4 minutes, inserting secondary travel activities (Koelet and Glorieux, 2006) and re-weighting of the *FTUS* data, the data re-combination procedure with the *FHTS* data was redone. The combined data were then utilized to simulate a travel survey data set for the 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.

The combined data are randomly split into two sets (training set: 75% and testing set: 25% of the data). In the first step towards setting up the simulation, the training set is used to categorize individuals into "homogeneous" groupings with respect to the respondents' duration of trips conducted per day conditional on the purpose of travel. In this paper, the work travel purpose is reported. The *CART* method, a computationally intensive exploratory classification tool proposed by Breiman *et al.*, (1984) is used to this effect. This method involves growing an overly large tree to capture all potentially important splits; then pruning the tree back to the root node to create a hierarchy of sub trees; and finally, selecting an optimal-sized tree from this sequence using an independent holdout sample or cross-validation. 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 independent variables used to explain the mobility indicator - the total travel duration per person (travel participant) per day, include: gender, age group, marital status and education level. Thus conditional on the respondents' travel purpose, "homogeneous" categories of the trip durations of individuals per day are developed using the *CART* method. 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 household members (Greaves, 2006).

The next step involves developing frequency distributions from which one can sample for the travel characteristic of interest. Within each of the established "homogeneous" categories developed in the preceding stage, the trip durations exhibit some variation. To capture this variation therefore, discrete frequency distributions of values of total trip durations per person per day were developed. The 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 a 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. Simulations are run 500 times. The simulated data is then compared with the reserved testing data using descriptive statistics so to examine the difference between the simulated and the real survey data.

3 Results and Discussion

In practice, *HTS* may cover only a small sample of the population thus failing to be representative of the target population. This makes the combination of *HTS* data with data from other related surveys, a vital solution so as to obtain larger data base samples that can be used for simulation. A set of general guidelines on what practitioners may consider when intending to perform data integration includes: Examining background information on travel data and other data sources; Reconciliations of concepts and definitions; Re-categorization, re-coding and transformation of variables; and harmonizing time periods of pre-integration data sets.

For the entire combined data set, as would be expected, the majority of the trips (about 38%) are conducted due to returning home, 13% due to work, also 13% attributed to shopping and 11% are sports/culture/relaxation-related trips. The rest of the travel is due to much less conducted trips for purposes of picking/dropping someone, obtaining services (doctor, bank), following education, visiting someone and making business visits. Focusing on the work travel purpose, 75% of the combined 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 CART resulted into 5 groupings. The most important variable on which the data are split is gender, followed by education and the age group of respondents. Marital status was not found to be important in explaining travel duration. Table 1 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 combined data), the testing set (25% of the combined data) and finally for the simulated data, in the last column.

Table 1: Comparison of the Durations of Work Trips per Participant per Day

No. in Training Data	No. in Testing Data	Categorization Scheme	Mean in minutes (Standard deviation)		
			Training Data	Testing Data	Simulated Data
640	219	Gender[1]=2	25.52(21.33)	25.49(23.92)	25.85(23.91)
55	23	Gender=1, Education[2]=1	21.76(18.05)	27.52(28.38)	23.51(23.79)
369	130	Gender=1, Education=2, 3, 4, Age group[3] =1, 3	32.31(33.56)	41.02(61.09)	32.62(35.36)
325	96	Gender=1, Education=2, 3, Age group=2	36.10(38.42)	34.57(51.27)	34.47(36.12)
187	51	Gender=1, Education=4, Age group=2	43.10(40.67)	51.99(59.89)	40.58(42.03)
1599	527	Overall	31.01(31.52)	33.53(45.56)	30.49(31.93)

[1]Codes 1, 2 are defined as male and female respectively; [2]Codes 1,2,3,4 are Primary school, Junior high school, High school and College or University respectively; [3]Codes 1,2,3 are '16-34', '35-54' and '55-75' years respectively.

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. Furthermore, males with primary education are found to travel for the shortest duration. It can be noticed that the travel duration of females tends to be homogeneous, irrespective of their age and education. This trend is also observed in the simulated data. The testing data set follows more or less the same trend but with a few deviations. This could be attributed to chance and smaller samples. It can also be noted that there is high variation between peoples' travel durations per day across all generated 'homogeneous' groups. Generally, the results show that the simulated travel duration data does not dramatically differ from the real survey data.

4 Conclusions

This paper has utilized travel survey data and time use data from Flanders, combined with population socio-demographic as the base data. Conditional on the travel purpose, the combined data were split into two sets. The training set was used to classify individuals into groups that exhibit similar ranges of travel durations. The generated groups were then used as the basis of the simulation. The reserved testing set was exploited in making comparisons with the simulated data. Overall, it becomes clear in this study that simulated travel data provides relatively accurate travel estimates. Simulated travel data can thus replicate real survey data and can be viewed as a good supplement of travel survey data. This approach serves as a powerful and practical tool in situations where travel data is lacking or inadequate. It further enables working with much larger samples. However, it is evident that any systematic biases or unusual behavior in the base travel data would also be reflected in the simulated travel data. If the existing data does not adequately capture the behavior of the representative sample, then, neither will the simulated data.

REFERENCES

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

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.

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.

Koelet, S., and I. Glorieux. Presentatie en Vergelijkbaarheid van de Onderzoeken TOR'99, NIS'99 en OVG'00 in het kader van Mobiliteitsonderzoek. (TOR 2006/24). Vrije Universiteit Brussel, Belgium, 2006 (in Dutch).

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.

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*a*) 'The Impact of Data Integration on Some Important Travel Behavior Indicators' Paper presented at the 86th Annual Meeting of the Transportation Research Board (TRB), and is to be published in the Transportation Research Record (TRR) of TRB, Washington DC, January 2007.

Nakamya, J., Moons, E. and Wets, G. (2007*b*) 'Combining Survey Data from Different Studies to Simulate a Local Travel Survey Sample Data Set: An Application to the Flemish Region. 'Paper to be presented at the World Conference on Transport Research (WCTR) in Berkeley, USA, June 2007.

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.