INVESTIGATION OF THE REQUIRED TRAVEL SURVEY SIZE FOR TRAINING AN ACTIVITY-BASED TRAFFIC DEMAND MODEL FOR FLANDERS IMPLEMENTED IN THE FEATHERS SIMULATION PLATFORM

dr. ir. Bruno Kochan
Prof. dr. ir. Tom Bellemans
Prof. dr. Davy Janssens
Prof. dr. Geert Wets (*)

dr. ir. B. Kochan, Prof. dr. ir. T. Bellemans, Prof. dr. D. Janssens and Prof. dr. G. Wets are members of the Transportation Research Institute (IMOB), Faculty of Applied Economics, Hasselt University
Wetenschapspark 5 Box 6
B-3590 Diepenbeek
Belgium
Fax. + 32(0)11 26 91 99

E-mail: Bruno.Kochan@UHasselt.be
Tel: +32(0)11 26 91 47
E-mail: Tom.Bellemans@UHasselt.be
Tel: +32(0)11 26 91 27
E-mail: Davy.Janssens@UHasselt.be
Tel: +32(0)11 26 91 28
E-mail: Geert.Wets@UHasselt.be (*) Corresponding author
Tel: +32(0)11 26 91 58

Abstract	222
Text	3859
Figures	4 * 250 = 1000
Tables	4 * 250 = 1000
Total Word Count :	**6081**

Date submitted: November 13, 2012

Abstract. It has been known from many years now that operational activity-based models need a lot of survey data to incorporate behavioural decision making of people. While there have been contributions from the field of statistics about how much survey data is needed to come to reliable estimates of behaviour; an obvious question which is often overlooked in the domain is how much survey data is really necessary to obtain an activity-based model that is sufficiently competent and accurate. This question is not only scientifically challenging and interesting, but also can significantly reduce data collection costs and is also very useful for practitioners. A very appealing question would be whether an activity-based model could also be trained with a smaller survey data set without losing too much model quality. This paper tries to explore this research question in the case of an activity-based model for Flanders (Belgium) inside the 'Forecasting Evolutionary Activity-Travel of Households and their Environmental RepercussionS' (FEATHERS) framework. As the scheduler in this study is based on decision trees, progressive sampling is being applied in order to investigate accuracy for all discrete choice decision trees. Based on the results of this investigation, it is demonstrated that for some decision trees the activity-based survey data set can be very small without losing accuracy, while for other decision trees bigger data sets are needed.

INTRODUCTION

Travel demand models are an important tool used in the transportation planning process to analyze transportation policies and decisions. Transportation forecasts have traditionally followed the sequential 4-step model where 4 sequential steps from trip generation to traffic assignment yield traffic volumes on network links. Activity-based models on the other hand form another class of transportation demand models that predict on an individual level where and when specific activities (e.g. work, leisure, shopping, etc.) are conducted. Along with these activities also trips are generated so that the end result of an activity-based model consists of activity-travel diaries or schedules. Activity-based models can be developed as standalone applications, however they can also be embedded in a framework that allows the models to be created, updated and maintained more easily. One such framework is the FEATHERS framework (*1*). The idea was conceived to develop a modular activity-based model of transport demand framework, where the emphasis is on the one hand on the practical use of the system by practitioners and end users and on the other hand on facilitating the creation of alternative activity-travel demand models. Similar initiatives, like for instance the Multi-Agent Transport SIMulation toolkit (MATSIM) (*2*) and the Common Modelling Framework (CMF) (*3*) have been developed for trip-based and complex tour-based models, highlighting the potential relevance of such a modular system. The activity-based scheduling model that is implemented in the FEATHERS framework and that is used in this study is based on the scheduling model that is present in A Learning BAsed TRansportation Oriented Simulation System (ALBATROSS) (*4*).

However, any activity-based model needs activity-based survey data in order to be trained. Typically, a subset of the entire study population is selected for participating into a travel survey. These travel surveys might take many different forms such as PAper and Pencil Interviews (PAPI), Computer Assisted Telephone Interviews (CATI), Web Assisted Personal Interviews (WAPI), Tablet Assisted Personal Interviews (TAPI), Personal Digital Assistant Interviews and others. All of these survey forms have their advantages and disadvantages, however most surveys being set up will cost a certain amount of money which restricts the number of respondents being interviewed as monetary resources are in most cases limited (*5*). Therefore, an interesting question is how small a travel survey size can be in order to still have an activity-based model with reliable estimates of the population's travel behaviour. This is what will be investigated in this study.

To date, partial and fully operational activity-based micro simulation systems include A Learning BAsed TRansportation Oriented Simulation System (ALBATROSS) (*6*), Micro-analytic Integrated Demographic Accounting System (MIDAS) (*7*), the Activity-Mobility Simulator (AMOS) (*8*), Prism Constrained Activity-Travel Simulator (PCATS) (*9*), Florida's Activity Mobility Simulator (FAMOS) (*10*) and other systems developed and applied to varying degrees in Portland, Oregon, San Francisco, Florida and New York.

As stated before, all these activity-based models need activity-travel survey data as input for training the model. These travel survey data set sizes differ from study area to study area. For example, for the ALBATROSS (*6*) model as it is employed in The Netherlands, the survey for that study invited households to fill out an activity diary for two consecutive days. This resulted in a total of 5.295 household records that were used for training the model. As will be further discussed in this paper, for the study presented in this paper, the travel survey for Flanders in total has 8.800 person records as only 1 person per household was surveyed.

This paper is organized as follows: first, a general overview of the FEATHERS activity-based framework is given, where after the currently available activity-based model based on the ALBATROSS core, which resides in FEATHERS, will be discussed. In a second part the travel survey data for this study is explored, so that in a third step the actual analysis

of the accuracy of a step-wise larger survey data set is explored based on an accuracy indicator. Subsequently, major conclusions are drawn and avenues for future research are worked out.

FEATHERS AND THE ALBATROSS SYSTEM

In order to facilitate the development and maintenance of dynamic activity-based models for transport demand, the FEATHERS framework was developed. For this purpose FEATHERS provides the tools needed in order to develop and maintain activity-based models in a particular study area. The framework supplies tailored memory structures such as, 'households', 'persons', 'activities', 'trips', 'cars', etc. and at the same time FEATHERS is also equipped with a database structure that is able to nourish activity-based models being developed, assimilated or modified inside FEATHERS. In such a way, in the framework, users can opt for a wide variety of functionalities that are provided by the FEATHERS modules facilitating the creation and maintenance of activity-travel demand models. Because of these properties, FEATHERS is very suitable for the research proposed in this paper.

Currently the FEATHERS framework incorporates the core of the ALBATROSS Activity-Based scheduler (*6*). This scheduler assumes a sequential decision process consisting of decision trees that intends to simulate the way individuals build schedules. The output of the model consisting of predicted activity schedules, describes for a given day which activities are conducted, at what time (start time), for how long (duration), where (location), and, if travelling is involved, the transport mode used and chaining of trips.

The underlying methodology and assumptions used in each major step within the ALBATROSS model are as follows. The scheduler first starts with an empty schedule or diary where after it will evaluate whether or not work activities will be included. If this is the case, then the number of work activities will be estimated together with their beginning times and durations. In a second step the locations of the work activities are determined. The system sequentially assigns locations to the work activities in order of schedule position. This is done by systematically consulting a fixed list of specific decision trees. During the third step the model proceeds with the next decision steps, that is: selection of work related transport modes, inclusion and time profiling of non-work fixed and flexible activities, determination of fixed and flexible activity locations and finally determination of fixed and flexible activity transport modes.

ACTIVITY-BASED TRAVEL SURVEY DATA USED IN THIS STUDY

Activity-based models differ highly from traditional transport forecasting models in the sense that the former models aim at predicting the interdependencies and interrelationships between the multitude of facets of activity profiles on an individual level. The major distinction with conventional models is that scheduling of activities comprises the foundation of activity-based models. Therefore, and in line with the basic underpinnings of the activity-based paradigm, the data required to estimate an activity-based model differs from the data required to build conventional models. More specifically, in order to build an activity-based model of transport demand, data on activity patterns are required. While there is a wide variety of possible types of travel surveys that can be employed for the purpose of estimating conventional transport models, the primary objective of the data collection effort for activity-based models should be reflective of the data necessary to estimate this kind of model. Current household travel surveys rely extensively on the use of mail, telephone, internet and multimedia methods to obtain information on the daily travel and other activities of a representative sample of the population. Given the needs of the activity-based modelling

approach, the travel survey to be called in has to pay attention on the measurement of activities at the end of trips and to how and when the respondent chose to do them. One such travel survey for the Flemish study area that can be used for estimating the activity-based model inside FEATHERS is the *Onderzoek VerplaatsingsGedrag Vlaanderen* (OVG) travel survey. This OVG survey formally is a trip-based survey method, however information about trip purposes and hence information about activities in between trips is available. Therefore, this survey is particularly suitable for estimating the activity-based model embedded in the FEATHERS framework. The OVG travel survey was conducted through 8.800 persons that were selected based on a random sample from the national register. These persons (1 randomly chosen adult per household) were all involved in a 1-day survey that was conducted primarily through face-to-face interviews. During these surveys information about the demographic, socioeconomic, household and trip-making characteristics of these individuals were gathered and for the purpose of this research, all person records and their according travel were then processed and being used as input for estimating the activity-based model incorporated inside FEATHERS.

INVESTIGATION OF THE MINIMUM SURVEY SAMPLE SIZE PER DECISION TREE

Having access to these travel survey data sets does not necessarily imply that induction algorithms for training decision trees must use them all. Smaller samples often provide the same accuracy. However, the correct sample size rarely is obvious. In this paper a method of progressive sampling is being applied where progressively larger samples of the original travel survey data are used as long as model accuracy improves till a point where no improvement can be obtained. In case smaller samples of the original travel survey data set can be used without losing model accuracy, then smaller surveys might be employed, and thus saving monetary costs and precious time as processing survey data can be very time-consuming.

Progressive sampling

The requirement for accurate models often demands the use of large data sets that allow algorithms to discover complex structures and make accurate parameter estimates. In this paper we study a progressive sampling method, which attempts to maximize accuracy. Progressive sampling starts with a small sample and uses progressively larger ones until model accuracy no longer improves. A central component of progressive sampling is a sampling sequence $S = \{n_0, n_1, n_2, ..., n_k\}$ where each n_i is an integer that specifies the size of a sample to be provided to an induction algorithm. In this study, the induction algorithm used in the activity-based model inside FEATHERS, is the Chi-squared Automatic Interaction Detector (CHAID) (*11*).

A learning curve (Figure 1) depicts the relationship between sample size and model accuracy. The horizontal axis represents n, the number of instances in a given training set, that can vary between zero and N, the total number of available instances. The vertical axis represents the accuracy of the model produced by an induction algorithm, in this case CHAID, when given a training set of size n.

Learning curves typically have a steeply sloping portion early in the curve, a more gently sloping middle portion, and a plateau late in the curve. The middle portion can be extremely large in some curves (*12, 13, 14*) and almost entirely missing in others. The plateau occurs when adding additional data instances does not improve accuracy. The plateau, and even the entire middle portion, can be missing from curves when N is not sufficiently large.

Conversely, the plateau region can constitute the majority of curves when N is very large. It is assumed that learning curves are well behaved. Specifically, it is assumed that the slope of a learning curve is monotonically non-increasing with n except for local variance. When a learning curve reaches its final plateau, we say it has converged. We denote the training set size at which convergence occurs as n_{min}. Given a data set, a sampling procedure, and an induction algorithm, n_{min} is the size of the smallest sufficient training set. Models built with smaller training sets have lower accuracy than models built from training sets of size n_{min}, and models built with larger training sets have no higher accuracy. Figure 1 shows an example sampling sequence and its relation to a learning curve. Empirical estimates are necessary to determine n_{min}. In general, these characteristics are not known in advance, thus, in many cases, n_{min} is nearly impossible to determine from theory. However n_{min} can be approximated empirically by a progressive sampling procedure.

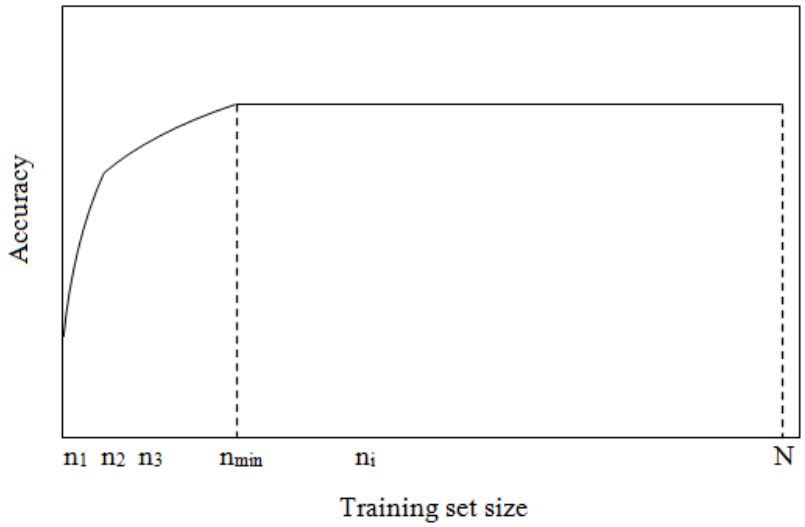

FIGURE 1 Learning curves and progressive samples.

Different kinds of progressive sampling approaches can be distinguished. An elementary sampling approach, called arithmetic sampling (*15*), uses the following sequence:

$$S_a = n_0 + (i.n_\delta) = \{n_0, n_0 + n_\delta, n_0 + 2.n_\delta, ..., n_0 + k.n_\delta\} \quad (1)$$

An example arithmetic sampling sequence is {100, 200, 300, ..., n_k}. An alternative sampling sequence approach is called geometric sampling and uses the following sequence:

$$S_g = a^i.n_0 = \{n_0, a.n_0, a^2.n_0, a^3.n_0, ..., a^k.n_0\} \quad (2)$$

In this paper, the authors have chosen to work with an arithmetic sampling approach, because of its simplicity and because of the fact that the data sets we are dealing with are not that large when compared with huge data sets that sometimes exist in the field of data mining.

In progressive sampling the learning curve is being evaluated. However this learning curve essentially is the chain of values of the accuracy indicator for each decision tree. Which accuracy indicator is being used in this study is explained in the section below.

The accuracy indicator

While performing progressive sampling, an accuracy indicator is being calculated for each discrete decision tree of the activity-based model. The accuracy indicator that is chosen in this study is the Confusion Matrix Accuracy (CMA) measure (*16*). If a classification system has been trained to distinguish between choices, a confusion matrix will summarize the results of testing the algorithm for further inspection. The CMA value of a decision tree is calculated by calculating a ratio where the nominator is calculated by the sum over all cells in the confusion matrix of the decision tree, and the denominator is calculated by the sum over all diagonal cells of the confusion matrix. The following example in table 1 will illustrate this calculation of the CMA value of a decision tree. Let's assume we are dealing with a decision tree that can predict the transport mode of a trip. Let's further assume we have three transport modes, namely car, public transport and bike. Table 1 shows, as a fictitious example, the confusion matrix for this decision tree.

TABLE 1 Example of a confusion matrix

		Predicted class		
		Car	Public transport	Bike
Actual class	Car	5	3	0
	Public transport	2	3	1
	Bike	0	2	11

In this confusion matrix, of the eight actual cars, the system predicted that three were public transport, and of the six public transports, it predicted that one was a bike and two were cars. We can see from the matrix that the system in question has trouble distinguishing between cars and public transport, but can make the distinction between bike and other types of transport mode pretty well. All correct guesses are located in the diagonal of the table, so it's easy to visually inspect the table for errors, as they will be represented by any non-zero values outside the diagonal. The CMA value in this case equals 70.3 %.

Detecting convergence

A key assumption behind all the progressive sampling procedures discussed above is that convergence can be detected accurately and efficiently. Convergence detection is fundamentally a statistical judgement. As explained before, the learning curve is modelled as sampling progresses. However, at this point an important remark has to be made. Each sample being taken from the original travel survey data set is unique, that is to say, the original travel survey data set can be used in order to obtain different samples of the same size. Indeed, starting from a data set of a certain size, one can obtain many different samples of the same size by randomly selecting instances of the original travel survey data set. Therefore, the authors suggest to compile 30 different samples of the same size for each step in the progressive sampling approach. Each of the different 30 samples will yield a different group of decision trees and therefore also different CMA values. Table 2 shows the average values of those CMA values while table 3 shows a summary of all normalised average CMA values. Normalisation of the CMA values is necessary in order to compare learning curves between themselves. The learning curve for each decision trees is made by chaining the averages of the CMA values for each sample step in the progressive sampling approach. These points on the learning curve are then used to estimate a tangential line, for each of the 30 samples, so that the slope of the tangential line is compared to zero. As explained before, learning curves, as shown in figure 1, have three regions of behaviour, a primary rise, a secondary rise, and a plateau. Once the plateau is reached, the tangential line is approximately zero. At this point

we have taken advantage of a common property of learning curves: the slope of the line tangent to the curve constantly decreases. Taking into account the explanation above, the authors of this paper suggest to use the following criterion as a stopping criterion where convergence is being reached. Convergence is reached when more than 90% of the 30 sample CMA values at a given point on the learning curve have a tangent smaller or equal than 0.25 degrees, which is almost zero degrees or flat.

TABLE 2 Average CMA values, for each progressive sampling step from 10% till 90% (expressed in percent)

Nr	Decision tree	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	Inclusion work	75,7	75,9	75,7	75,7	75,7	75,8	75,8	75,8	75,7
2	Number of episodes	66,2	66,7	66,4	66,7	67,2	67,1	67,3	67,7	68
3	Location, same as previous	56,7	58,2	62,2	63,1	64,3	65,3	65,8	66,6	66,9
4	Location, in/out home	59,4	60,6	61,6	61,8	62,5	62,9	63,0	63,4	63,8
5	Location, order (1)	26,6	27,8	28,1	28,6	29,1	29,4	29,8	30,0	30,2
6	Location, nearest of order	69,0	71,1	71,7	72,3	72,3	73,0	73,5	73,8	74,2
7	Location, distance band (1)	20,4	21,0	22,5	23,2	23,8	24,5	25,0	25,5	25,5
8	Location, order (2)	31,2	32,6	33,6	34,0	34,4	35,1	35,6	36,0	36,4
9	Location, distance band (2)	29,9	32,3	34,0	35,6	36,4	36,9	37,4	37,6	38,0
10	Transport mode (1)	58,8	60,6	60,9	61,5	61,9	62,8	63,0	63,4	64,4
11	Inclusion fixed	86,7	87,1	87,1	87,2	87,2	87,2	87,3	87,3	87,3
12	Number of episodes	45,2	46,3	47,1	47,0	47,1	47,4	47,6	47,9	48,1
13	Chaining, work	46,2	47,1	47,6	48,3	48,9	49,3	49,7	49,9	50,3
14	Location, same as previous	58,4	58,4	59,8	60,2	60,3	60,4	60,4	60,4	60,5
15	Location, distance-size class	6,3	6,9	7,4	7,8	8,1	8,3	8,3	8,4	8,4
16	Inclusion flexible	78,8	78,9	79,0	79,1	79,2	79,2	79,3	79,3	79,4
17	Duration	37,0	37,8	38,4	38,8	39,1	39,3	39,4	39,5	39,6
18	Timing	39,7	43,2	44,7	45,6	46,1	46,2	46,1	46,2	46,2
19	Chaining	82,2	86,2	86,7	87,1	87,5	87,4	87,4	87,5	87,4
20	Transport mode (2)	47,8	50,3	51,8	52,7	53,3	53,7	53,8	53,9	53,8

Table 4 shows for all discrete decision trees the convergence of the accompanying learning curves. When taking a look at table 1 some interesting conclusions can be drawn. First of all, there are 3 decision tree learning curves that reach convergence immediately at a 10% fraction of the total travel survey data set. Astonishingly they constitute the inclusion decision trees. Apparently the inclusion decision trees do not need much travel survey data in order to be accurate enough. Figure 2 shows an example of the learning curve of the first decision tree which models the inclusion of work activities. As can be clearly seen, the learning curve is flat, meaning that the 3th region of the learning curve, the plateau, is reached almost immediately. Another interesting conclusion that can be drawn is that most other decision trees have a learning curve that reach the plateau when the sample fraction is between 10% and 90 %. Figure 3, showing the learning curve of decision tree 18, gives an example. The last conclusion that can be made is that there are a few decision trees that do not reach convergence and hence there is no plateau. This means that for these decision trees more travel survey data is needed in order to have decision trees with a fine accuracy. Not surprisingly it comprehend decision trees that are used to make location choices. This is in line with expectation, as location choices by nature are more difficult to make when compared with decisions regarding the inclusion of a certain activity like for instance a work activity.

TABLE 3 Normalised average CMA values, for each progressive sampling step from 10% till 90% (expressed in percent)

Nr	Decision tree	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	Inclusion work	100,1	100,3	100,0	100,0	100,0	100,1	100,1	100,1	100,0
2	Number of episodes	97,0	97,7	97,3	97,8	98,4	98,3	98,5	99,1	99,6
3	Location, same as previous	84,3	86,5	92,4	93,8	95,5	97,1	97,8	98,9	99,5
4	Location, in/out home	92,9	94,6	96,2	96,5	97,6	98,2	98,4	99,1	99,6
5	Location, order (1)	87,0	90,7	91,7	93,5	95,2	96,0	97,5	97,9	98,6
6	Location, nearest of order	92,2	95,0	95,7	96,6	96,5	97,5	98,1	98,5	99,1
7	Location, distance band (1)	77,1	79,4	85,1	87,5	89,7	92,4	94,4	96,1	96,5
8	Location, order (2)	85,5	89,2	92,0	93,1	94,0	96,2	97,6	98,6	99,7
9	Location, distance band (2)	78,1	84,6	89,0	93,3	95,3	96,6	97,8	98,4	99,5
10	Transport mode (1)	90,6	93,3	93,8	94,8	95,3	96,8	97,1	97,7	99,3
11	Inclusion fixed	99,2	99,7	99,7	99,8	99,8	99,8	99,9	99,9	100,0
12	Number of episodes	94,4	96,7	98,3	98,2	98,3	98,9	99,4	100,0	100,5
13	Chaining, work	92,5	94,4	95,5	96,8	98,1	98,9	99,6	100,0	100,9
14	Location, same as previous	96,9	97,0	99,2	99,9	100,1	100,2	100,3	100,3	100,4
15	Location, distance-size class	75,7	83,0	89,5	93,8	97,5	100,4	100,6	101,2	101,6
16	Inclusion flexible	99,2	99,3	99,4	99,5	99,6	99,7	99,8	99,8	99,9
17	Duration	93,8	95,9	97,3	98,3	99,2	99,7	99,8	100,0	100,3
18	Timing	85,7	93,3	96,4	98,3	99,3	99,6	99,5	99,7	99,7
19	Chaining	94,2	98,7	99,3	99,7	100,2	100,1	100,1	100,2	100,1
20	Transport mode (2)	88,5	93,1	95,8	97,5	98,6	99,4	99,7	99,7	99,7

TABLE 4 Percentage of the number of samples with a tangent smaller than 0.25 degrees, for each progressive sampling step from 10% till 90% (expressed in precent)

Nr	Decision tree	10%	20%	30%	40%	50%	60%	70%	80%	90%
1	Inclusion work	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
2	Number of episodes	90,0	100,0	97,0	97,7	100,0	100,0	97,7	100,0	100,0
3	Location, same as previous	60,0	13,3	53,3	70,0	63,3	77,7	86,7	96,7	100,0
4	Location, in/out home	73,3	80,0	100,0	96,7	93,3	100,0	100,0	100,0	100,0
5	Location, order (1)	40,0	60,0	63,3	63,3	80,0	70,0	93,3	86,7	76,7
6	Location, nearest of order	50,0	86,7	93,3	100,0	100,0	100,0	100,0	100,0	100,0
7	Location, distance band (1)	60,0	10,0	50,0	56,7	50,0	56,7	53,3	76,7	30,0
8	Location, order (2)	36,7	43,3	80,0	83,3	63,3	63,3	96,7	96,7	100,0
9	Location, distance band (2)	20,0	16,7	23,3	73,3	73,3	90,0	100,0	86,7	100,0
10	Transport mode (1)	50,0	86,7	90,0	100,0	76,7	100,0	96,7	93,3	100,0
11	Inclusion fixed	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
12	Number of episodes	53,3	56,7	96,7	96,7	90,0	93,3	96,7	93,3	100,0
13	Chaining, work	53,3	66,7	70,0	76,7	76,7	96,7	96,7	86,7	100,0
14	Location, same as previous	90,0	63,3	100,0	100,0	100,0	100,0	100,	100,0	100,0
15	Location, distance-size class	6,7	16,7	23,3	26,7	40,0	93,3	86,7	83,3	100,0
16	Inclusion flexible	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
17	Duration	60,0	90,0	93,3	96,7	100,0	96,7	100,0	100,0	100,0
18	Timing	3,3	26,7	76,7	96,7	100,0	100,0	100,0	100,0	100,0
19	Chaining	10,0	93,3	100,0	100,0	100,0	100,0	100,0	100,0	100,0
20	Transport mode (2)	16,7	40,0	63,3	90,0	90,0	100,0	100,0	100,0	100,0

Figure 4 shows, as an example, the learning curve of the first location decision tree that does not reach a plateau.

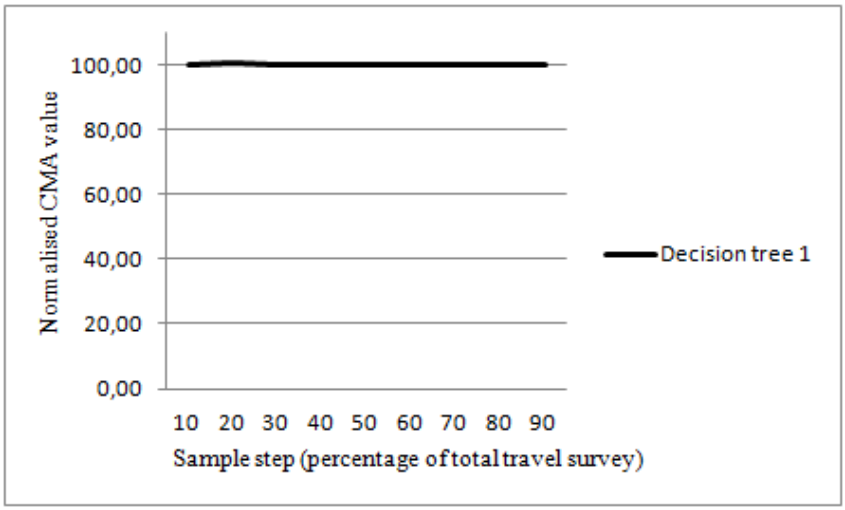

FIGURE 2 Learning curve of decision tree 1.

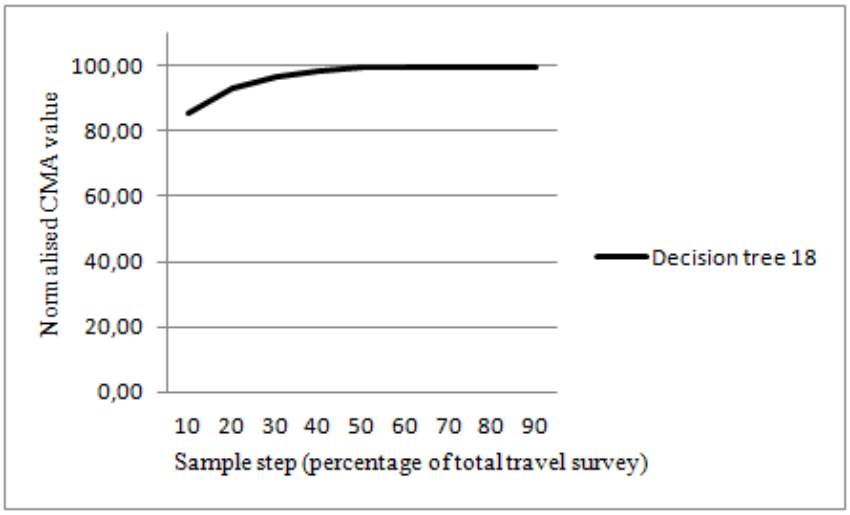

FIGURE 3 Learning curve of decision tree 18.

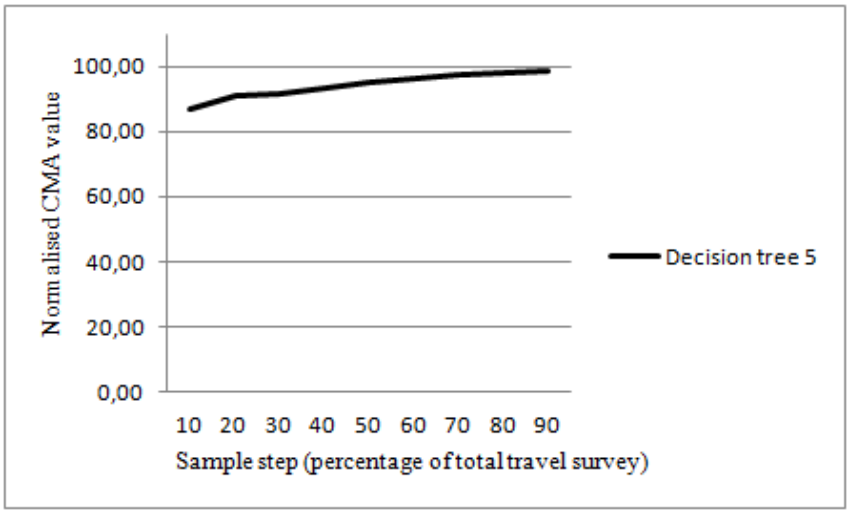

FIGURE 4 Learning curve of decision tree 5.

CONCLUSIONS AND DISCUSSION

The study performed in this paper tried to find out if it is possible to train an activity-based model, based on decision trees, with a smaller travel survey data set than the original one used for the study area Flanders. As it appeared, for some decision trees, namely the inclusion decision trees, extreme small survey data sets can be used, data sets that are of a size 10% of the original survey data set. For the majority of the decision trees smaller travel data sets could be used without losing model accuracy, while for a few decision trees, all location-related decision trees, more survey data should be used in order to have a model with high accuracy. Overall, the conclusion is that in general, all decision trees taken together, it is not advisable to work with small travel survey data sets as there are a few decision trees that are strongly affected by the survey size. Making use of a small travel survey would decrease the accuracy of an activity-based model, especially the location model which is an important component, as activity-based models in transportation research are used for predicting trips. These trips and hence origin-destination matrices could become less reliable when the model is trained with only a small travel survey data set.

In this paper we only focused on the decision trees constituting the activity-based model components. However, it could be interesting to deduce decision trees based on increasingly smaller travel survey data sets and then based on those new decision trees to predict travel behaviour of the Flemish population in order to investigate whether or not there is an impact of the smaller decision trees on travel behaviour in Flanders. This could be done for future research.

REFERENCES

(1) Bellemans T., B. Kochan, D. Janssens, G. Wets, T. Arentze, and H. J. P. Timmermans. Implementation Framework and Development Trajectory of the Feathers Activity-Based Simulation Platform. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2175, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 111-119.

(2) MATSIM, Multi Agent Traffic SIMulation. http://www.matsim.org. Accessed July 12, 2012.

(3) Davidson, W., R. Donnelly, P. Vovsha, J. Freedman, S. Ruegg, J. Hicks, J. Castiglione, and R. Picado. Synthesis of First Practices and Operational Research Approaches in Activity-Based Travel Demand Modelling. *Transportation Research Part A*, Vol. 41, 2007, pp. 464-488.

(4) Arentze T., and H. J. P. Timmermans. The sensitivity of Activity-Based Models of Travel Demand: Results in the Case of ALBATROSS, In A. Jaszkiewicz, M. Kacmarek, J. Zak, M. Kubiak, Advanced OR and AI methods in Transportation, 2005, pp. 573-578.

(5) Rosero-Bixby L., J. Hidalgo-Céspedes, and D. Antich-Montero. Improving the Quality and Lowering Costs of Household Survey Data Using Personal Digital Assistants (PDAs). An Application for Costa Rica. In 2005 meeting of the Population Association of America, Philadelphia, March 31, 2005.

(6) Arentze T. and H. J. P. Timmermans. *ALBATROSS 2: A Learning-Based Transportation Oriented Simulation System*, European Institute of Retailing and Services Studies, The Netherlands, Eindhoven, 2005.

(7) Goulias K. and R. Kitamura. *A dynamic model system for regional travel demand forecasting, in Panels for Transportation Planning: Methods and Applications*, Kluwer Academic Publishers, U.S.A., Boston, pp. 321-348, 1996

(8) Kitamura R., E.I. Pas, C.V. Lula, T.K. Lawton, and P.E. Benson. The Sequenced Activity Mobility Simulator: An Integrated Approach to Modelling Transportation, Land Use and Air Quality. In *Transportation*, No. 23, 1996, pp. 267-291.

(9) Kitamura R. and S. Fujii. *Two Computational Process Models of Activity-Travel Behavior*. *Theoretical Foundations of Travel Choice Modelling*, Elsevier Science, Oxford, pp. 251-279, 1998

(10) Pendyala R. M. Phased Implementation of a Multimodal Activity Based Modelling System for Florida. FAMOS: The Florida Activity Mobility Simulator. Final Report Submitted to the Florida Department of Transportation, Research Centre. Volume I: Technical Documentation Available at: http://www.dot.state.fl.us/research-center/Completed_Proj/Summary_PTO/FDOT_BA496rpt.pdf and Volume II: Users Guide Available at: http://www.dot.state.fl.us/research-center/Completed_Proj/Summary_PTO/FDOT_BA496_Manual.pdf. Accessed July 12, 2012

(11) Kass G.V. An Exploratory Technique for Investigating Large Quantities of Categorical Data. In *Applied Statistics*, Vol. 29, 1980, pp. 119-127.

(12) Catlett J. Megainduction: a test flight. In proceedings of the 8th International Workshop on Machine Learning, Morgan Kaufmann, pp. 596-599, 1991.

(13) Catlett J. Megainduction: Machine learning on very large databases. PhD thesis, School of Computer Science, University of Technology, Australia, Sydney, 1991.

(14) Harris-Jones C., and T.L. Haines, Sample size and misclassification: Is more always better ? Working paper AMSCAT-WP-97-118, AMS Centre for Advanced Technologies, 1997.

(15) John G., and Langley P. Static versus dynamic sampling for data mining. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 367-370, 1996.

(16) Kohavi, R., and F. Provost. *Glossary of terms: Machine Learning*. Kluwer Academic Publishers, Boston, 1998.