The Development of a Segmentation Scheme for the Evaluation and Prediction of Activity-Travel Sequences

Davy Janssens
Geert Wets[1]
Tom Brijs
Koen Vanhoof

Hasselt University - Campus Diepenbeek
Transportation Research Institute
Wetenschapspark 5, bus 6
BE - 3590 Diepenbeek, Belgium
Tel: +32(0)11 26 {91 28; 91 58; 91 55; 9153}
Fax: +32(0)11 26 91 99
E-mail: {davy.janssens; geert.wets; tom.brijs; koen.vanhoof}@uhasselt.be

Word count	
Abstract:	247
Text:	5928
Figures:	4*250
Tables:	1*250
Total number of words:	7425

Submission date: 27/07/06

[1] Corresponding author

ABSTRACT

In this paper, sequential information in data is represented and captured through the use of Markov Chains. The core knowledge information in a Markov Chain is typically represented by means of transition matrices, revealing information about the underlying structure of the data sequence. A drawback of the current application of Markov Chains is that there is only one transition probability matrix which is both representative for every person (respondent) and for every time frame during the day. To this end, a novel segmentation procedure has been introduced and tested in this paper that enables one to cluster transition matrices in terms of time and socio-demographic information. The temporal segmentation used the technique of the identification of bifurcation points; the socio-demographic segmentation used a modified version of a decision tree, in the sense that sequential probability information was used during induction and in the leaves of the tree as opposed to the traditional way of only using one single classification attribute. The segmentation procedures were both adopted for descriptive and predictive purposes in the empirical section. Results show that the technique reveals promising information both at the descriptive and predictive level. At the descriptive level, evidence was found that one should rely upon different transition probability matrices for different time windows during the day and that socio-demographic information should be taken into account as well. For prediction purposes, the segmentation approaches simulated more accurate activity-travel sequences at pattern level; while the opposite was found true at trip level.

1. INTRODUCTION AND PROBLEM FORMULATION

Within the field of activity-based transport modelling, the activity agenda and the corresponding travel participation is typically represented by means of sequences or patterns of behaviour. Sequential data representation is also frequently adopted in research on trip chains, activity sequencing, and sequential choice of activities and locations, acknowledging that consecutive activities likely affect one another (*1-6*). Accordingly, it is clear that a sequential data representation seems to be the relevant unit of analysis in the activity-based transportation research domain (see also (*7*)).

A sequence can be defined as a succession of events, while an event is a transition from one discrete state to another, situated along a time continuum (*8*). In this paper, events represent activities that occur in a persons' diary. Traveling is considered as an activity as well, while transport mode is added as an additional attribute in this case. In order to analyze sequential data, the techniques and methodologies need to be chosen most carefully.

In this paper, we will represent and capture sequential information in data through the use of Markov Chains. Markov Chains are typically represented by means of transition matrices, which reveal information about the underlying structure of the data sequence and provide information about the probability of going from the previous activity (state) to the current activity (state) (see also section 2).

While Markov chains have been frequently adopted in several domains for analysis and prediction, the technique is certainly not applied on a broad scale in the area of transportation. However, also in other research domains it is –unfortunately- often assumed in the application of Markov Chains that there is only one transition probability matrix which is both representative for every person (respondent) and for every time frame during the day (which is called the stationarity assumption). However, especially in the field of transportation, there is accumulated empirical evidence which suggests that activity-travel patterns are (highly) correlated with the socio-demographic information (*9-10*) of the respondent and that different transport behaviour (and thus different activity-travel patterns) exist for different time windows during the day (*11-12*).

The aim of this paper is therefore to propose advancements to the current methodological state-of-the-art for application within the field of transportation research and activity-based modelling. The first advancement is the introduction of a segmentation procedure which enables one to cluster transition matrices in terms of time information (relaxation of the stationarity condition) by means of a technique which is able to identify statistically significant bifurcation points. As a second advancement, a new segmentation scheme has been developed which is able to cluster sequential information in terms of socio-demographic information. This procedure uses a modified version of a decision tree, in the sense that sequential probability information can be used during induction and in the final nodes in the tree (leaf nodes) as opposed to the traditional way of only using one single classification attribute (represented by one dependent variable).

The remainder of this paper has been organised as follows. Section 2 briefly gives an introduction into Markov Chains and transition matrices and explains how a test for stationarity can be adopted in this context. Next, a new segmentation approach has been introduced which is both able to take into account time and socio-demographic segmentation. The fourth section explains both descriptive and predictive empirical results using the information incorporated in the segmentation approaches. The final section gives a conclusion and discusses some topics for future research.

2. MARKOV CHAINS

2.1 Transition Matrices

Markov Chains are probabilistic models which were introduced by Andrej Andreevic Markov at the beginning of the 20th century. Their application domains have been numerous; for a more comprehensive treatments of Markov Chains and their applications see e.g. (*13*). In addition to these applications, there is also a large family of latent variable models that can be jointly used along with Markov Chains. They can for instance be used as Mixed Markov Latent Class models (MMLC) (see examples in (*14-15*)) that have been used to describe stochastic processes in discrete space and discrete time. Finally, Markov Chains are also particularly well suited in the analysis of longitudinal (panel) data.

A transition matrix reveals information about the underlying structure of the data sequence and is in fact the core knowledge representation of a Markov Chain. In order to capture sequential dependencies in activity-travel data and represent it in terms of a transition matrix, it is assumed that each sequence in the diary consists of a set of correlated successive observations of a random variable. To this end, a discrete random variable X_t is considered, taking values in the finite set $\{1,..m\}$, where each value in this set represents an activity that occurs in a persons diary. As mentioned above, travelling is considered as an activity as well, however the transport mode is added as an additional attribute in this case. The goal in this application is to generate (predict) the value taken of X_t as a function of the values taken by previous observations of this variable. On the one hand, one can assume that the current value taken by X_t can be entirely explained by the previous observation (Activity *t-1*). On the other hand, one can assume that it is only possible to accurately explain the current value of X_t by the last *k-1* observations (Activity *t-1*, Activity *t-2*, … Activity *k-1*) (i.e. *k-1*th lag) in which *k* represents the length of the diary.

While we have explained in previous studies (*16*) the importance for choosing the number of previous observations that can best explain the current observation in the diary, the trade-off is not dealt with in this paper. The reason is that we wanted to investigate the methodological advancements that have been proposed in this paper in a simple setting where we assume that the current value of X_t can be entirely explained by the previous observation. In this case, it is in fact implicitly assumed that:

$$P(X_t=i_0 \mid X_0=i_t, \ldots, X_{t-1}=i_1) = P(X_t=i_0 \mid X_{t-1}=i_1) = q_{i_1 i_0}(t), \text{ where } i_t, \ldots, i_0 \in \{1,\ldots, m\}.$$

Each value in the set $\{1,\ldots,m\}$ represents an activity that occurs in a persons diary. Considering all combinations of i_1 and i_0, we can now construct a transition probability matrix Q, each of whose rows sums to 1.

$$Q = \begin{array}{c|cccc} X_{t-1} \backslash X_t & 1 & \cdots & \cdots & m \\ \hline 1 & q_{11} & \cdots & \cdots & q_{1m} \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ m & q_{m1} & \cdots & \cdots & q_{mm} \end{array}$$

Each of these transition probabilities $q_{..}$ represents the probability of going from state X_{t-1} to state X_t, or in other words, from the activity$_{t-1}$ to activity$_t$. Alternatively, probabilities within the matrix can also be represented in terms of frequencies, which is often referred to as a transition frequency matrix Q_{freq}.

2.2 Testing for Stationarity

A Markov Chain satisfies the stationarity condition if transition probabilities do not depend on the time *t*. It means that at whatever time point *t* the chain is looked at, transition probabilities are the same.

Stationarity can be tested by means of an "omnibus" method (*17*) that divides a sequence (in this case activity-travel patterns) into *D* time periods, thereby yielding *D* subsequences. Then, transitional frequency matrices are computed for each time period (*d*) and a statistical chi-square test will compare the individual transition frequency matrices with the overall one. The expected transitional frequencies and the χ^2-statistic can be respectively computed as:

$$E_{ij}(d) = n_{itot}(d)\frac{n_{ij}}{n_{itot}} \forall i, j$$

$$\chi^2 = \sum_{d=1}^{D} \sum_{i=1, j=1}^{m} \frac{(n_{ij}(d) - E_{ij}(d))^2}{E_{ij}(d)}$$

The tested null-hypothesis is that the transitional probabilities are constant across time periods. It is expressed as:

H_0: $q_{ij}(d)=q_{ij}$, $\forall$ d=1,2,…, D

H_1: $q_{ij}(d)\neq q_{ij}$, $\forall$ d=1,2,…, D.

3. THE NEED FOR A NEW SEGMENTATION APPROACH

As mentioned before in the introduction, empirical evidence seems to suggest that activity-travel patterns are (highly) correlated with the socio-demographic information (*9-10*) of the respondent and that the stationarity condition described above is often violated (*11-12*). When dealing with time segmentation, only one explanatory variable (i.e. time of day) needs to be taken into account. Therefore, the statistical test of stationarity of a system that has been introduced in the previous section, which mainly examines the change of dynamics before and after a certain moment in time, can be extended by considering different splitting points that lead to a significant statistical difference. Things get more complicated when segmentation needs to be done in terms of socio-demographic information, because of different explanatory variables which might have an influence on the final segmentation. In this case, a modified version of a decision tree has been developed, such that the dependent variable in the tree explicitly takes sequential information per socio-demographic variable into account. Both approaches have been described in the next two sections.

3.1 Segmentation by Means of Bifurcation Points (Temporal Segmentation)

The stationarity test in section 2.2. has introduced a statistical test for examining whether there is a change of dynamics before and after a particular cutpoint. For instance, the test can be used to examine whether transition matrices are significantly different in the time periods ranging from 24PM-8AM; from 8AM-16PM and from 16PM-24PM. Obviously, the test can equally be used for a segmentation in more segments. The choice of these cutpoints can be done arbitrarily, relying for instance on domain knowledge. While this is a good procedure to get an initial idea about whether our transition matrices satisfy the stationarity assumption or not, splitting a sequence at different single points in time is unlikely to result in the most optimal segmentation because it is not at all driven by

the information which is incorporated in the data. Therefore, by the iterative application of the procedure described in section 2.2 (omnibus test), for all possible splitting points of a sequence, we are able to point out moments in time where the system bifurcated into significantly different type of dynamics. Such a procedure is also valid to evaluate whether the identified pivotal moments in the data match with the moments defined by a priori domain knowledge. The points that radically transform the dynamics of a system are called bifurcation points. The methodology is frequently used in complex dynamical system theory (see for instance (*18*)). The procedure to identify these bifurcation points is as follows:

1. Determine the level of significance (α).
2. Set a time window by which transition matrices need to be compared. In the limit, this time window can be set equal to 1 minute but this will lead to a computational explosion of the calculations. In the artificial example given above (24PM-8AM; 8AM-16PM and 16PM-24PM), the three time windows are set equal to 8 hours, and the potential bifurcation points are set at 8AM, 16PM and 24PM. Accordingly, every time window defines potential bifurcation points (n_1,n_2,n_3). The first potential bifurcation point is defined as n_{min}, the last as n_{max}.
3. Construct a transition matrix for every time window, meaning three transition matrices in this example.
4. Calculate the χ^2 value to evaluate whether the dynamics of the system is subject to a segmentation into *d* (i.e. 3 by example) time periods (see definition 5.5) by application of the omnibus test.
5. Store the *p*-value for this omnibus test.
6. Redefine the time window ranging from n_{min} to n_{max} by adding one time window to n_{min}, thereby setting n_{min}:=n_{min}+1. Recalculate the transition matrices for the new time windows. In our example, a new transition matrix needs to be computed ranging from 24PM-16PM. The transition matrix that was computed before, for 16PM-24PM, remains the same.
7. Re-calculate the omnibus test for n_{min} to n_{max}. Equally, substract one time period from *d*; i.e. *d*:=*d*-1. In our example, it is thus evaluated whether the dynamics of the system is subject to a segmentation in two time periods.
8. Store the *p*-value for this omnibus test.
9. Repeat steps 6 till 8 until n_{min}:=n_{max} or until d:=1.
10. Plot the *p*-values for every omnibus test

The procedure will later be empirically illustrated in section 4.3.

3.2 Segmentation by means of Full Decision Trees (Socio-Demographic Segmentation)

3.2.1. Conceptualisation

A different and more complicated procedure arises when transition probability matrices need to be segmented in terms of socio-demographic information. Indeed, unlike in the previous case, there is now a combination of different explanatory variables that have a potential influence on the transition probability matrix. To this end, a novel segmentation scheme has been developed that is a modified version of a decision tree approach. Especially CART decision trees were used in a number of previous studies (*9*; *19*) in the context of transportation modelling for segmentation. The best known application of this technique is probably the TRANSIMS project, where the CART algorithm is used in the “Activity Generator Module” to produce an accurate classification of household characteristics based on household travel behaviours.

However, in traditional (classification) decision trees, the dependent variable at the leaf (a leaf node is a node that have no offspring nodes) simply contains a finite number of possible values and is often discrete in nature. The novel algorithm that is proposed in this paper differs in two ways from this common way of thinking. First, the dependent variable can no longer be immediately observed from the data but is the result of a learning methodology (i.e. transition matrices are extracted from the data, see section 2.1) and second, the dependent variable explicitly takes sequential information into account. As such, transition probability matrices instead of simple discrete values are used as dependent variables in the construction of the trees.

Obviously, the most important decision that needs to be made when developing a decision tree is the splitting criterion (this is the criterion which is used to divide the tree into several branches). Fortunately, one of the most widely measures that is adopted in decision trees, i.e. gain ratio, can be applied in a quite straightforward manner in our approach as well. The use of gain ratio as a split criterion favours splits into increasingly homogeneous partitions in terms of the dependent variable (class attribute), because the best split is the one with the most homogeneous daughters. In the limit, leaf nodes will therefore only contain cases from a single response class. Gain ratio is a measure which is derived from information theory. Information theory defines the quantity of information conveyed by a particular message as being inversely proportional to the predictability of that message. When a message is entirely certain (that is, its probability is 1), then the quantity of information conveyed is zero. When a message is nearly improbable (that is, its probability is almost 0), a maximum quantity of information is needed to receive such a message. The degree of uncertainty of a message can be represented by the probability of that message, or in terms of traditional decision trees, by the probability of that class. Information theory works with measures like entropy as an intermediary step in its computation to finally arrive at the gain ratio. Entropy can be defined as a measure for impurity, disorder and randomness of a particular system and the goal is typically to reduce the entropy in the system. Entropy is measured in bits. To summarize, in the case of decision tree induction, the aim is to reduce the entropy in the tree by recursively splitting the tree in the most homogeneous way. For a better understanding of the concepts of information theory, the reader may consider the work by Quinlan (*20*) in the context of traditional decision tree induction.

3.2.2. Re-introducing Information Theory in the Context of Transition Matrices

In order to calculate the gain ratio described above, traditional information theory needs to be tuned and re-introduced for application in the (new) context of transition matrices. Five steps can be distinguished in this process.

Step 1: Calculate the Entropy of One Row in the Matrix

The entropy of one row i in a frequency matrix Q_{freq} can be defined as:

$$\text{Info } Q_{freq}(i) = -\sum_{j=1}^{m} \left(\frac{(Q_{freq}(i,j)}{n_{itot}} \right) \times \log_2 \left(\frac{(Q_{freq}(i,j)}{n_{itot}} \right) \text{bits}, \forall \text{ row } i$$

, with $Q_{freq}(i)$ the i^{th} row of the frequency matrix Q_{freq}; $Q_{freq}(i, j)$ the matrix entries defined by the i^{th} row and the j^{th} column, that is the frequency that the element(s) in row i is (are) followed by the element in column j; n_{itot} the row total in Q_{freq} and m the number of columns.

Example: Assume that we have the following transition frequency matrix:

Q_{freq}=

	X_t				
X_{t-1}	T_c	E	F	R	n_{itot}
T_c	5	18	2	5	30
E	1.7	3.68	4.09	0.53	10
F	6.25	10	8.75	5	30
R	0	1.7	5.8	2.5	10

,with T_c= **T**ransportation, with **c**ar as transport mode, F=visit **F**amily, E=**E**at, R=**R**ead

The entropy of row two is then:

$$InfoQ_{freq}(2)= -\frac{1,7}{10}\log_2(\frac{1,7}{10}) - \frac{3,68}{10}\log_2(\frac{3,68}{10}) - \frac{4,09}{10}\log_2(\frac{4,09}{10}) - \frac{0,53}{10}\log_2(\frac{0,53}{10}) = 1,72 \text{ bits}$$

The calculation of the entropy of other rows in the matrix is similar.

Step 2: Calculate the Entropy of the Total Matrix

Second, the entropy of a full transition matrix is equal to $Info(Q)=\sum_{i=1}^{m}\frac{n_{itot}}{N}\times Info(Q_{freq}(i))$. It can be seen from this formula that every row in the transition matrix is weighted in proportion to the number of times a particular sequence starts with the element(s) which is represented in that particular row of the matrix.

Example: The entropy of the full transition probability matrix, shown in the example above is equal to: (30/80)x1,56+(10/80)x1,72+(30/80)x1,95+(10/80)x1,38 =1,70 bits.

Step 3: Calculate the Entropy of a Transition Matrix After a Partition on Test X

Third, the entropy of a full transition probability matrix after a (sub)set has been partitioned on a test *X* using a decision tree for instance, can be calculated as:

$Info_X(Q)=\sum_{i=1}^{n}\frac{|T_i|}{|T|}\times info(Q)$, where $|T_i|$ represents the number of cases that belongs to the partition *i* and $|T|$ represents the number of cases in *T*.

Example: Assume that the first branch is specified by the transition probability matrix introduced previously (see step 1), and that a second branch contains a transition probability matrix that has a total entropy equal to 1,50 bits. Assume now that both branches respectively represent 4 and 3 cases for this particular split (*X*). The calculation is as follows: (4/7)x1,70+(3/7)x1,50=1,61 bits.

Step 4: Calculate the Gain Criterion

Fourth, one needs to calculate the gain criterion, which measures the information that is gained by partitioning a set using a particular test *X* in a decision tree. The gain criterion can be defined as: Gain (*X*)= info (*T*)-info$_X$(*T*).

When compared to traditional decision tree induction, its application does not change when transition matrices are used in the leaves of the tree. The calculation is straightforward by using the formulas explained in steps 1 till 3.

Step 5: Calculate the Gain Ratio

It was already mentioned before that the previous steps were mainly intermediary in order to arrive at a measure called the gain ratio. The gain ratio can be calculated as Gain ratio $(X)=\frac{gain(X)}{split\ info\ (X)}$, where split info ($X$) indicates the information that is generated by partitioning T into n subsets. It is calculated by:

$$-\sum_{i=1}^{n}\frac{|T_i|}{|T|}\times\log_2\left(\frac{|T_i|}{|T|}\right).$$

The gain ratio is preferred over the gain criterion for splitting a decision tree, since attributes that have a large number of possible values, give rise to a multiway branch with many child nodes, when information gain criterion is used for the calculation. More information about this property can be found in Quinlan (*20*). Also in this case, the calculation of gain ratio does not change in this renewed context.

3.3 The Segmentation Procedure

3.3.1. A New Decision Tree Segmentation Scheme

In order to use these modified principles of information theory in a new decision tree segmentation scheme, the following mathematical conceptualisation has been introduced.

S the total sample of activity diaries, consisting of n sequences, indexed $i=1,\ldots,n$

X_k explanatory socio-demographic attributes, with $k=1,...,K$.

Y dependent variable, represents the transition probability matrix Q ($Y=Q$); for all sequences ($\forall\ i$)

T Final decision tree based on sequential information, comprised of nodes (*Ns* and *L*) and branches. Leaf nodes are specified by Q.

N the current node in T, splitting the current subset of S into subsets N_{kt}

N_{kt} represents the subsets of a split by N; splits at a value t based on an independent variable X_k, such that $X_k=t$; $\forall t_k$

t_k set of possible values of t such that there exist observations in N having $X_k=t$; $\forall k=1,\ldots,K$; and with $N=X_k$.

Ns set of active decision nodes in T that split S into different subsets.

L set of inactive decision nodes that cannot split S into additional subsets because n_{min} or G_{min} are not satisfied (see infra). In this case they become leaf nodes L.

n_{min} Parameter that determines whether a particular branch in the tree is split into additional nodes or not. Splitting is stopped when the number of individuals that belong to either of the child nodes N_{kt} is less than the number defined by n_{min}

$G(N_{kt})$ Gain ratio (as defined in section 3.2.2) of a transition matrix that is built on the subset of N_{kt}.

Max (GN_{kt}) represents the global maximum of all the gain ratios per level in the tree, **$G(N_{kt})$** $\forall N_{kt}$. Max $G(N_{kt})$ is used to select the optimal decision node *N*.

G_{min} Minimum Gain ratio that determines whether a particular branch in the tree is split into additional decision nodes or not. Splitting is stopped when Max $G(N_{kt})$ is smaller than G_{min}.

Having defined this mathematical conceptualisation, a decision tree procedure has been introduced in Figure 1. A computer code has been established to automate the full process. The next paragraph elaborates on an example to illustrate this procedure.

3.3.2. Example

Consider the following 3 activity diaries (activity-travel pattern) for illustration purposes:
Diary 1: T_cEEFREREERFT_cFT_cFFT_cFET_cF
with gender=male; age=older than 45 years and education=high
Diary 2: RREFEFEE$T_c$$T_c$R
with gender=female, age=between 18 and 24 and education=low
Diary 3: EEFFT_cFT_cFRR$T_c$$T_cRT_c$RR
with gender=male, age=between 25 and 44 and education=low
and with T_c= **T**ransportation, with **c**ar as transport mode, F=visit **F**amily, E=**E**at, R=**R**ead

The initialisation procedure in Figure 1 is quite simple for this example. The value n_{min} and G_{min} are respectively set equal to 1 and 0, X_k is defined as gender, age and education, respectively for k=1,…,3; with K=3. The set of active decision nodes *Ns* is also fixed as {gender, age, education}. Note that this set is not always equal to the variables X_k for k=1,…,K, since one might decide not to use certain variables as decision nodes, for example in case of ID-number, which might be perfectly relevant as an attribute but not as a decision node. Finally, the set of leaf nodes is initialised as empty.

Since the set of decision nodes *Ns* is not empty and the set of leaf nodes is empty, the two first checks in Figure 1 can be omitted. After this, the procedure will select the most optimal decision node for the root node of the tree.

First, each decision node *N* that belongs to *Ns*, is divided into temporary splits N_{kt} such that X_k=t $\forall N_{kt}$ for k=1,…,3. Then, $Q_{N_{kt}}$ is constructed for each t_k. Next, the gain ratio is calculated as explained in section 3.2.2, and the attribute that achieves the highest gain ratio in the tree is selected to carry out the split at the current level (=root level) of the tree. The attribute "Gender", with a gain ratio of 0.469, achieves the highest value (Max$G(N_{kt})$) and is thus chosen as the best split for this tree at root level. After actually creating this split, the next step first verifies whether the number of observations in the child nodes, is greater than the minimal value n_{min} and greater than G_{min}. While this is the case for the first branch (gender=male), it is not for the second branch (gender=female). For this reason the decision node "Gender" is removed from the set of active decision nodes *Ns* and added to the set of leaf nodes. The branch for which the decision node did not satisfy the n_{min}-check (i.e. N_2) is marked to indicate that it has been fully exploited. The procedure now restarts from the beginning. While the first check still is not yet satisfied, the set of leaf nodes is no longer empty. First, the set of active decision nodes is temporarily set equal to the set decision nodes X_k. This is necessary to let a particular variable occur multiple times in different branches in the tree. Second, the most left unmarked branch in this tree is now identified. In this example this unmarked branch is specified by Gender=male. Again, the most optimal *N* needs to be determined for this branch. This means that temporary splits need to be created for the branch gender=male and the maximum gain ratio need to be computed. In this case, the variable education achieves the largest gain ratio. Now, both remaining

branches (i.e. education=low; education=high) do not satisfy the n_{min}-check. Indeed, for the branch gender=male; there is only one case that belongs to education=low and one case that belongs to education=high. This means that both branches need to be marked. While *Ns* is still not empty, all branches are now marked and the final decision tree along with its final *Q* are stored and shown in Figure 2.

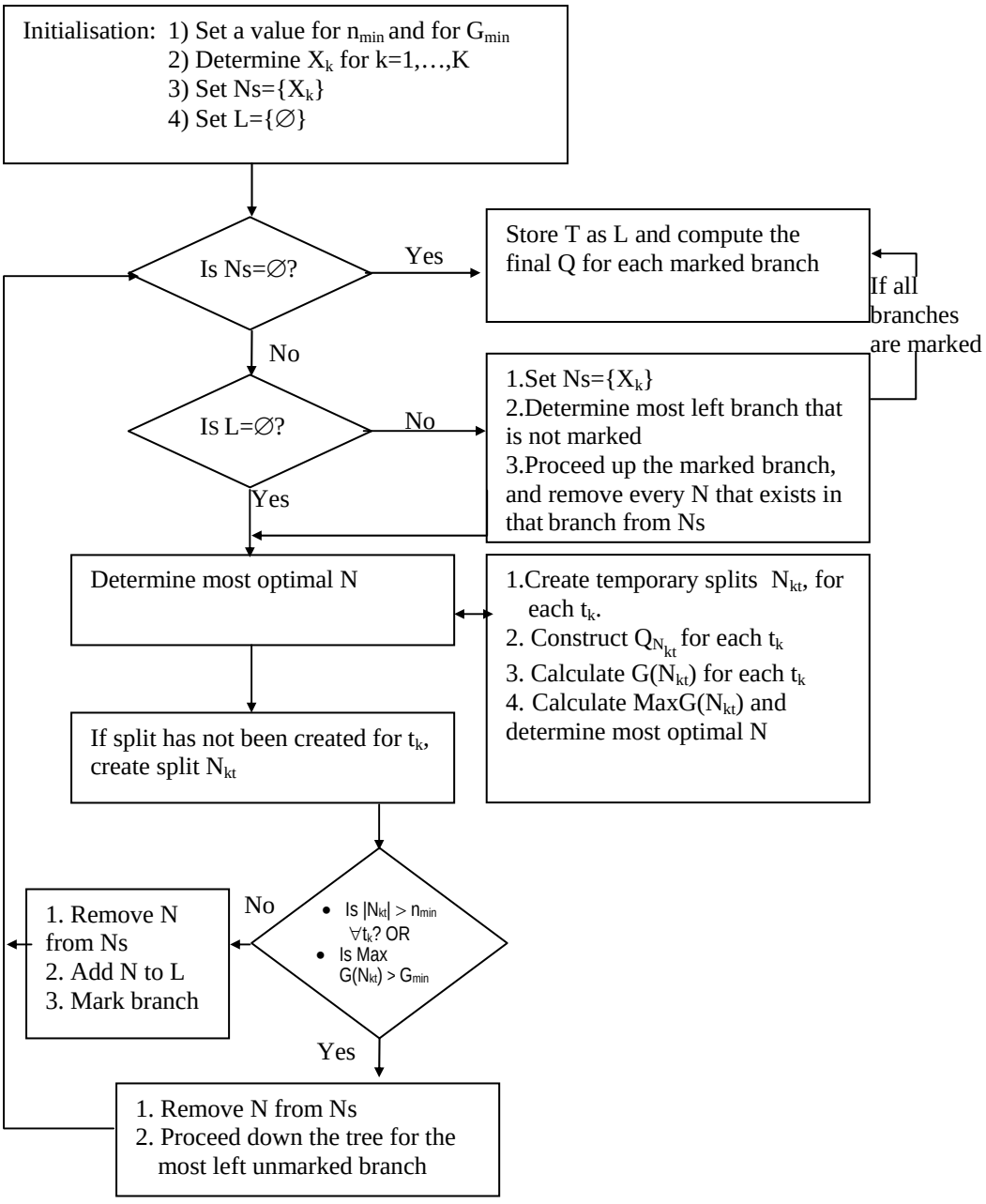
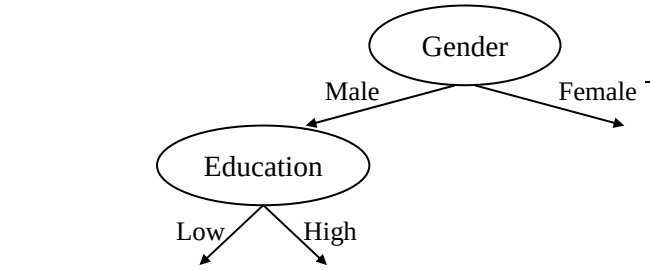


FIGURE 1 Description of the procedure for building a decision tree based on sequential information.



	T_c	F	E	R
T_c	0,00	0,80	0,20	0,00
F	0,50	0,17	0,17	0,17
E	0,17	0,17	0,33	0,33
R	0,00	0,33	0,67	0,00

	T_c	F	E	R
T_c	0,5	0	0	0,5
F	0	0	1	0
E	0,25	0,5	0,25	0
R	0	0	0,5	0,5

	T_c	F	E	R
T_c	0	1	0	0
F	0	0,94	0,06	0
E	0	0	0	0
R	0	0	0	0

FIGURE 2 The final sequential information decision tree (example).

Once the segmentation for all the transition probability matrices has been done, every sequence can use the most appropriate transition matrix by proceeding its socio-demographic information down the tree.

4. EMPIRICAL RESULTS

4.1. Preface

The previous sections have described segmentation procedures which enable one to cluster transition matrices in terms of time and socio-demographic information. Obviously, the aim of these procedures is twofold. First, there is an analytical purpose, where one is able to determine whether different segments exist in terms of socio-demographic and temporal information in the activity-travel pattern data. Second, if these segmentations can be found, they can be adopted in a simulation scheme to evaluate their impact when used for prediction purposes. We will illustrate both purposes in this empirical section. With respect to the simulation scheme that has been used, the procedure that has been introduced in (*16*) will be used in our experiments. The core of this simulation procedure is that it predicts the value X_t based on the previous lag, being X_{t-1} and then the next value to be predicted becomes X_{t+1}, which is based on X_t (predicted in previous step). This repetition continues until the simulated activity diary equals the length of the diary, simulated in the beginning of the procedure. For more details, we refer to Janssens *et al.* (*16*).

4.2 Data

The activity diary data used in this study were collected in the municipalities of Hendrik-Ido-Ambacht and Zwijndrecht in the Netherlands (South Rotterdam region) to develop the Albatross model system (*21*). The data involve a full activity diary, implying that both in-home and out-of-home activities were reported. The sample covered all seven days of the week, but individual respondents were requested to complete the diaries for two designated consecutive days. Respondents were asked, for each successive activity, to provide information about the nature of the activity, the day, start and end time, the location where the activity took place, the transport mode, the travel time, accompanying

individuals and whether the activity was planned or not. A pre-coded scheme was used for activity reporting. After cleaning, a data set of a random sample of 1649 respondents was used in the experiments. Household and person characteristics have been used which might be relevant for the segmentation of the sample.

4.3. Temporal Segmentation (Bifurcation Points)

The iterative application of the omnibus test is a computationally demanding procedure. In order to reduce the computational burden, we have defined time periods of 60 minutes. Consequently, there are 24 potential bifurcation points in the beginning of the procedure. As explained before in section 3.1, time windows will gradually be combined together, ending up with two time windows in the end, and every time defining new potential bifurcation points. The level of significance in our experiments was set at 5%. An evolution of the p-values is shown in Figure 3. The first p-value in the figure is the result of a comparison between 24 time windows (i.e. one transition matrix for every hour in the day). While there are obviously large (significant) differences between transition matrices that are built during morning periods (e.g. 3AM-4AM) and noon periods (e.g. 12AM-13PM), the differences are non-significant when the full day is considered. The reason for this is that during the majority of the day (except for some specific time periods), activity-travel combinations are more or less randomly distributed and majority patterns flatten out the significant differences. In other words, the dynamics of the system do not change (alter) significantly every hour. In order to determine a more significant change in dynamics, more aggregated time periods need to be considered. For this reason, it may be somewhat surprising that time periods dividing the diary in for instance 8 time periods (i.e. every three hours) were found not significantly different during one day. Also in this case, while there are significant differences between the time periods 3AM-6AM and 12AM-3PM; the majority of the three-hour during time periods appeared to be non-significant. The significant effect starts to appear from 5 time periods (i.e. every 4 hours and 48 minutes) and ranges till 3 time periods (every 6 hours). Surprisingly, two time periods (lasting 12 hours each) were found not significantly different. One possible explanation is that the (frequently occurring) work-, sleep- and travel-combinations appear fairly equal in both time windows.

Since the stationarity condition has been violated for at least some time windows, it has now been experimentally determined that we should rely upon different transition probability matrices when predicting activity-travel combinations for these different time windows during the day.

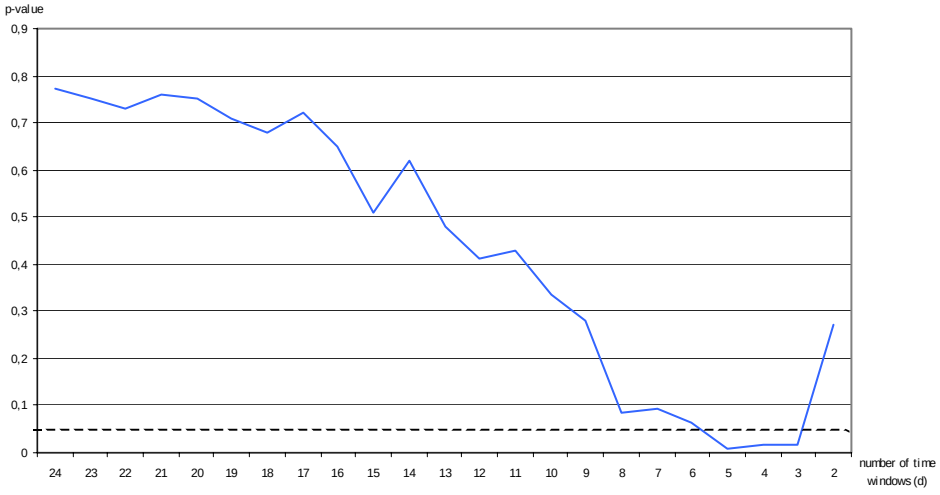


FIGURE 3 Evolution of p-values for the procedure described in section 3.1.

When we want to take this information into account in the simulation procedure to predict the activity-travel combinations, we will rely upon the first finding that was found significant, being 5 time periods, defined as 3AM – 7:48AM; 7:48AM-12:36PM; 12:36PM-17:24PM; 17:24PM-22:12PM and 22:12PM-3AM.

4.4. Socio-Demographic Segmentation

Segmenting transition matrices in terms of socio-demographic variables, assumes the execution of the procedure that was explained in section 3.3.1. The n_{min} parameter and the minimum gain ratio (G_{min}) were respectively arbitrarily set at 75 cases and at 0.05 to prevent overfitting of the tree on the training data. The final decision tree that was built for our data is shown in Figure 4. In addition to the structure of the decision tree, every decision node shows the number of cases that go down that branch, the maximum gain ratio and the information value that was achieved. Every leaf node,

```
(start)
| NCAR=1 (900, INFO = 2.727906, G = 0.058563)
| | HHTYPE=3 (311, INFO = 2.724312, G = 0.123689)
| | | GENDER=1 (309, INFO = 2.725375, G = 0.057060)
| | | | SEC=3 (134, INFO = 2.625469, G = 0.101251)
| | | | | BIKEAV=1 (124, INFO = 2.604755, G = 0.125868)
| | | | | | CARAV=1 (117, INFO = 2.602145, G = 0.098740)
| | | | | | | CHILDREN=2 (L)
| | | | | | | CHILDREN=1 (L)
| | | | | | | CHILDREN=3 (L)
| | | | | | | CHILDREN=4 (L)
| | | | | | CARAV=unknown (L)
| | | | | BIKEAV=unknown (L)
| | | | SEC=1 (L)
| | | | SEC=unknown (L)
| | | | SEC=2 (96, INFO = 2.684212, G = 0.128843)
| | | | | BIKEAV=1 (83, INFO = 2.650120, G = 0.130249)
| | | | | | CHILDREN=2 (L)
| | | | | | CHILDREN=1 (L)
| | | | | | CHILDREN=3 (L)
| | | | | | CHILDREN=4 (L)
| | | | | BIKEAV=unknown (L)
| | | | SEC=4 (L)
| | | GENDER=2 (L)
| | HHTYPE=1 (L)
| | HHTYPE=4 (239, INFO = 2.697180, G = 0.074627)
| | | GENDER=1 (212, INFO = 2.680736, G = 0.080723)
| | | | AGE=2 (153, INFO = 2.663894, G = 0.089576)
| | | | | SEC=3 (L)
| | | | | SEC=1 (L)
| | | | | SEC=unknown (L)
| | | | | SEC=2 (L)
| | | | | SEC=4 (L)
| | | | AGE=3 (L)
| | | | AGE=4 (L)
| | | | AGE=1 (L)
| | | GENDER=2 (L)
| | HHTYPE=5 (140, INFO = 2.651551, G = 0.114153)
| | | GENDER=1 (132, INFO = 2.650076, G = 0.095268)
| | | | SEC=3 (L)
| | | | SEC=1 (L)
| | | | SEC=unknown (L)
| | | | SEC=2 (L)
| | | | SEC=4 (L)
| | | GENDER=2 (L)
| | HHTYPE=2 (160, INFO = 2.560276, G = 0.139788)
```

```
(continued)
| | | CARAV=1 (154, INFO = 2.550203,
                G = 0.105228)
| | | | NBIKES=4 (L)
| | | | NBIKES=1 (76, INFO = 2.505884,
                  G = 0.125667)
| | | | | AGE=2 (L)
| | | | | AGE=3 (L)
| | | | | AGE=4 (L)
| | | | | AGE=1 (L)
| | | | NBIKES=unknown (L)
| | | | NBIKES=2 (L)
| | | | NBIKES=3 (L)
| | | | NBIKES=0 (L)
| | | CARAV=unknown (L)
| | HHTYPE=unknown (L)
| NCAR=unknown (164, INFO = 2.627780,
                G = 0.087061)
| | SEC=3 (L)
| | SEC=1 (L)
| | SEC=unknown (L)
| | SEC=2 (L)
| | SEC=4 (L)
| NCAR=2 (290, INFO = 2.656737,
          G = 0.062742)
| | GENDER=1 (267, INFO = 2.649541,
              G = 0.061261)
| | | SEC=3 (91, INFO = 2.520677,
                G = 0.109946)
| | | | CHILDREN=2 (L)
| | | | CHILDREN=1 (L)
| | | | CHILDREN=3 (L)
| | | | CHILDREN=4 (L)
| | | SEC=1 (L)
| | | SEC=unknown (L)
| | | SEC=2 (L)
| | | SEC=4 (131, INFO = 2.639051,
                G = 0.094076)
| | | | CHILDREN=2 (L)
| | | | CHILDREN=1 (L)
| | | | CHILDREN=3 (L)
| | | | CHILDREN=4 (L)
| | GENDER=2 (L)
| NCAR=5 (L)
| NCAR=0 (L)
| NCAR=3 (L)
| NCAR=4 (L) (end)
```

FIGURE 4 A final sequential information decision tree (empirical data).

containing different transition probability matrices was indicated by (L). It can be seen from this tree that the variable "number of cars" ("Ncar") was the most important variable in the tree, followed by Household type ("Hhtype"), gender ("Gender") and socio-economic class ("Sec"). Having applied temporal and socio-demographic segmentation, the full "knowledge model" is finalized and we are now ready to move on to the simulation of new activity-travel patterns.

4.5. Simulation Results

In order to prepare the simulation, data were divided into a training and a test set, thereby using the training set for building the model (transition matrices and segmentation tree), while the unseen test data were used for validation. Activity-travel patterns were simulated both for the training and the test data. The goodness-of-fit for the simulated diaries was measured by comparing the generated activity patterns with the observed patterns in the training and the test dataset. The comparison was measured using the following two indicators:

- Pattern level attributes (number of tours)
- Trip level attributes (trip rates)

Pattern level attributes give an indication about the performance of the simulation procedure at the highest level. While other indicators at pattern level might be chosen, the evaluation was made at this level by comparing the *mean* number of tours in the observed and the generated patterns and this for the training and the test set. The results of the simulated training set give an indication about how well the framework is capable of capturing and simulating the information which is incorporated in the training data. The decline in goodness-of-fit between this training set and the test set is taken as an indicator of the degree of overfitting. In case no segmentation has been used, this means that only one general transition matrix has been used for the prediction of the training and test data. In case segmentation is used, the combination of both socio-demographic and temporal information has been considered.

It can be clearly seen from Table 1 that when segmentation (temporal and socio-demographic segmentation) is taken into account, much better results could be achieved at pattern level and this both for the training and the test set. These results imply that the information incorporated in the transition matrices in the different branches of the tree, is a better representation than the use of one single transition matrix at pattern level.

Trip level attributes are lower in hierarchy, which means that not the whole pattern but the individual trip is taken as the relevant unit of analysis in the evaluation. Typically, trips are differentiated here by means of the main purpose for which the trip is undertaken. The mean trip rate is used as an evaluation measure. The mean trip rate is defined as the mean number of trips that a person has done during one particular day. It can be seen from Table 1 that the result which was found at pattern level –resulting in more accurate prediction for the segmentation solution– could not be maintained at trip level, for the test set. However, for the training set, segmentation still receives more accurate results. The reason for this seems obvious. Due to the fact that mean trip rates compare sequences at a more detailed level (individual trips are compared instead of large patterns); the segmentation solution (containing a large number of transition matrices) apparently had a negative impact on the overall simulation outcome due to an amount of overfitting which occurred at this level. The presence of overfitting in the case of segmentation can be seen in Table 1, where a high goodness-of-fit on the training data, is followed by a rather low goodness-of-fit on the test data. The same could not be observed in the case of no segmentation, resulting in lower degrees of overfitting on the training data and in a higher goodness-of-fit measure on the test set.

TABLE 1 Performance Evaluation between Observed and Predicted Sequences at Pattern and Trip Level

	Observed		Predicted			
			Without segmentation		With segmentation	
	Training set	Test set	Training set	Test set	Training set	Test set
Pattern level	2.801	2.435	1.722	1.621	2.732	2.232
Trip level						
• Work	0.738	0.735	0.702	0.744	0.731	0.823
• SL	0.572	0.569	0.536	0.597	0.560	0.471
• Service	0.491	0.496	0.471	0.502	0.482	0.431
• B/G	0.276	0.274	0.269	0.284	0.281	0.236
• Other	0.132	0.134	0.121	0.147	0.124	0.153

5. CONCLUSION

In this paper, sequential data which is highly present in activity-travel diaries have been represented by means transition probabilities which are stored in transition matrices. Those matrices can be considered as the core knowledge representation of a Markov Chain. While the technique is certainly not applied yet on a broad scale in the area of transportation, it is shown in the paper that it deserves further investigation and that the results are promising both for analytical and for prediction purposes.

A drawback of the current application of the technique is that there is only one transition probability matrix which is both representative for every person (respondent) and for every time frame during the day. To this end, a novel segmentation procedure has been introduced that enables one to cluster transition matrices in terms of time and socio-demographic information. The first segmentation used the technique of the identification of bifurcation points; the latter used a modified version of a decision tree, in the sense that sequential probability information was used during induction and in the leaves of the tree as opposed to the traditional way of only using one single classification attribute.

In the experiments, the segmentation was both realised for descriptive and predictive purposes. In the descriptive evaluation, it was found that for the temporal segmentation, the stationarity condition has been violated for at least some time windows, which means that one should rely upon different transition probability matrices for different time windows during the day. For the socio-demographic segmentation, a full decision tree has been constructed, resulting in the variables “number of cars” (“Ncar”), Household type (“Hhtype”), gender (“Gender”) and socio-economic class (“Sec”) being the most important variables. Both segmentations were used simultaneously in the analyses in a predictive simulation. At pattern level, the results are that the segmentation information incorporated in the transition matrices of the different branches in the tree, is a better representation than the use of one single transition matrix for prediction purposes. The same conclusion could not be found at trip level for the test set, which may be explained by the fact that segmentation, resulting in several transition matrices, seem to have a negative impact on the overall simulation outcome at trip level due to an amount of overfitting which seems to happen for this large number of transition matrices.

Further research should be conducted in order to get a better idea about the relative performance of the techniques that have been advanced in this paper against other clustering techniques that can be used to complement Markov Chains such as for instance the MMLC application in Goulias (*15*) and the Latent Class clustering application in Kim and Goulias (*22*).

Acknowledgement: Davy Janssens acknowledges support as a post-doctoral research fellow from the Research Foundation - Flanders (F.W.O.-Vlaanderen).

6. REFERENCES

1. Kitamura, R. Incorporating trip chaining into analysis of destination choice. *Transportation Research B*, Vol.18, 1984, pp. 67–81.
2. Hatcher, S.G. and H.S. Mahmassani. Daily variability of route and trip scheduling decisions for the evening commute. In *Transportation Research Record*: *Journal of the Transportation Research Board, No. 1357*, TRB, National Research Council, Washington, D.C., 1992, pp. 72–81.
3. Arentze, T., A. Borgers, and H.J.P. Timmermans. A model of multi-purpose shopping trip behavior. *Papers in Regional Science*, Vol. 72, 1993, pp. 239–256.
4. Kitamura, R., E.I. Pas, C.V. Lula, T.K. Lawton and P.E. Benson. The sequenced activity mobility simulator (SAMS): an integrated approach to modeling transportation, land use and air quality. *Transportation*, Vol. 23, 1996, pp. 267–291.
5. Timmermans, H.J.P. A stated choice model of sequential mode and destination choice behavior for shopping trips. *Environment and Planning A*, Vol. 28, 1996, pp. 173–184.
6. McNally, M.G. An activity-based micro-simulation model for travel demand forecasting. In Ettema, D.F., Timmermans, H.J.P. (Eds.), *Activity-Based Approaches to Travel Analysis*. Pergamon Press, Oxford, 1997, pp. 37-54.
7. McNally, M.G. (2000) The activity-based approach. Center for Activity Systems Analysis, Paper UCI-ITS-AS-WP-00-4.
8. Abbott, A. Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, Vol. 21, 1995, pp. 93-113.
9. Greaves, S.P. and P.R. Stopher. Creating a synthetic household travel and activity survey: rationale and feasibility analysis. In *Transportation Research Record*: *Journal of the Transportation Research Board, No. 1706*, TRB, National Research Council, Washington, D.C., 2000, pp. 82 - 91.
10. Veldhuisen, K., H.J.P. Timmermans and L.L. Kapoen. Ramblas: a regional planning model based on the micro-simulation of daily activity travel patterns. *Environment and Planning A*, Vol. 32, 2000, pp. 427-443.
11. Bhat, C.R. and S. Singh. A comprehensive daily activity-travel generation model system for workers. *Transportation Research A*, Vol. 34, 2000, pp. 1-22.
12. Hamed, M.M. and F.L. Mannering. Modeling travelers' postwork activity involvement: Toward a new methodology. *Transportation Science*, Vol. 27, 1993, pp. 381-394.
13. Doob, J.L. *Stochastic Processes*. John Wiley and Sons, Inc., New York, 1990.
14. Langeheine, R. and F. van de Pol. A unifying framework for Markov modelling in discrete space and discrete time. *Sociological Methods and Research*, Vol. 18(4), 1990, pp. 416-441.
15. Goulias, K.G. Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models. *Transportation Research B*, Vol. 33(8), 1999, pp. 535-557.
16. Janssens, D., G. Wets, T. Brijs, and K. Vanhoof. The development of an adapted Markov chain modelling heuristic and simulation framework in the context of transportation research. *Expert systems with applications*, Vol. 28, 2005, pp. 105-117.

17. Gottman, J.M., and A.K. Roy. *Sequential Analysis. A Guide for Behavioral Researchers.* Cambridge University Press, New York, 1990.

18. Lemay, P. *The Statistical Analysis of Dynamics and Complexity in Psychology: a Configural Approach*. Ph.D. Dissertation, Social and Political Sciences, University of Lausanne, Lausanne, 1999.

19. Vaughn, K.M., P. Speckman, and D. Sun. Identifying relevant socio-demographics for distinguishing household activity-travel Patterns: A multivariate regression tree approach. Paper prepared for The National Institute of Statistical Sciences (NISS), 1999.

20. Quinlan, J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, 1993.

21. Arentze, T.A. and H.J.P. Timmermans. *Albatross: A Learning-Based Transportation Oriented Simulation System*. European Institute of Retailing and Services Studies, Eindhoven, 2000.

22. Kim, T. and K.G. Goulias. Dynamic analysis of time use and frequency of activity and travel using Latent Class Clustering and Structural Equation Modeling. Paper presented at the Conference on Progress in Activity-Based Analysis, Maastricht, The Netherlands, 2004.