SIMULATING THE EFFECTS OF LIFE TRAJECTORY DECISIONS ON TRANSPORT MODE CHOICE: VALIDATION RESULTS

Marloes Verhoeven
Faculty of Architecture, Building and Planning - Eindhoven University of Technology - P.O. Box 513 - 5600MB Eindhoven - The Netherlands - G.G.Manzato@tue.nl

Theo Arentze
Faculty of Architecture, Building and Planning - Eindhoven University of Technology - P.O. Box 513 - 5600MB Eindhoven - The Netherlands - T.A.Arentze@tue.nl

Harry Timmermans
Faculty of Architecture, Building and Planning - Eindhoven University of Technology - P.O. Box 513 - 5600MB Eindhoven - The Netherlands - H.J.P.Timmermans@tue.nl

Peter van der Waerden
Faculty of Architecture, Building and Planning - Eindhoven University of Technology - P.O. Box 513 - 5600MB Eindhoven - The Netherlands – P.J.H.J.v.d.Waerden@tue.nl

ABSTRACT

The study of the dynamics of activity patterns across different time horizons is high on the research agenda in activity-based analysis. Mostly qualitative research has indicated that life course events such as moving house or a new job may trigger or force individuals and households to reconsider their activity-travel behaviour, leading sometimes to changes. Attempts of modelling this process in transportation research are still scarce however. Of one the few exceptions is Beige (2008), who has estimated a hazard model to that effect. In our previous research, we have explored the feasibility of developing Bayesian networks to capture the direct and indirect effects of life trajectory events on the dynamics of activity-travel patterns in general and transport mode choice in particular (Verhoeven et al., 2005, 2006, 2007). The present paper will further develop this line of work and discuss the results of some validation tests. In particular, we will examine the ability of a Bayesian belief network, learned from retrospective survey data, to successfully simulate (i) the occurrence of life course events; (ii) the interval times between events, (iii) the sequence of one-dimensional and multi-dimensional careers embedded in life trajectories and (iv) observed transport mode choice.

Keywords: Life trajectory, activity-travel behaviour, Bayesian belief networks

INTRODUCTION

The majority of studies in travel behaviour research are cross-sectional in nature. Choice probabilities are analyzed and modelled as a function of socio-demographics and the attributes of the choice options, using single observations for the sample at one point in time. Although such research is of interest, it does not provide any understanding of the dynamics that made people decide to change their behaviour. Recently, therefore, the dynamics of activity-travel patterns has received increasing attention. Studies have been concerned with multiple time horizons, ranging from long-term perspectives to short-run adaptation in the implementation of planned activity-travel agendas.

To contribute to the literature on long-term dynamics, in our previous research, we have explored the feasibility of developing Bayesian networks to capture the direct and indirect effects of life trajectory events on the dynamics of activity-travel patterns in general and transport mode choice in particular (Verhoeven et al., 2005, 2006, 2007). The work is based on the contention that life course events such as change of jobs, change of residential location, changes in household composition etc. may lead individuals to reconsider their current travel behaviour and change their behaviour if they feel that the change induced by such life course events will increase too much the pressure in coping with the changed circumstances. These life course events are represented as the nodes in a Bayesian belief network. Using relevant nodes, the structure of the network which represents and specifies the direct and indirect relationships between the life course events and aspects of activity-travel behaviour can then be learned the data.

In previous publications, we have presented this approach. The present paper will further develop this line of work and discuss the results of some validation tests. In particular, we will examine the ability of a Bayesian belief network, learned from retrospective survey data, to successfully simulate (i) the occurrence of life course events; (ii) the interval times between events, (iii) the sequence of one-dimensional and multi-dimensional careers embedded in life trajectories and (iv) observed transport mode choice.

The remainder of the paper is organized as follows. First, we will briefly describe the Bayesian beliefs networks that were learned from the data. This will be followed by the discussion of a series of validation tests. The paper is completed with conclusions.

THE NETWORKS

The Bayesian belief networks were learned, using the Hugin-PC algorithm (Andersen et al. 1989) and data of a retrospective survey held in 2004 involving 700 respondents. These respondents were asked to indicate the occurance and timing of a pre-defined set of life course events. In total, the life trajectories resulted in 7649 cases, where the unit of observation is a year and a person. In the survey, only the current transport mode choice was measured. Two networks were derived: one for the lifcycle events only and a second one, also included transport mode choice as an aspect of activity-travel behavior.

Two types of variables are included in the life trajectory network: (1) Personal Characteristics, such as Gender and Age, and (2) variables related to Structural Life course Events. The variable of interest is Occurrence Event. A year is chosen as unit for these models. This refers to one year of a persons life trajectory. In addition, external variables were included. The life course events are defined with respect to a certain moment in time, *t*, a certain person and type of event (using some classification of events) for which the model intends to predict whether or not a change occurred as a particular instance of an event of that type. The node Occurrence Event defines whether there was a change at time *t* for that certain event and what kind of change it was. The existing state before time *t*, e.g., residential situation in case of a housing event, is represented by the node labelled State. Time Ago (A) and Time Ago (B) define the time elapsed since the last occurrence of certain types of changes A and B, respectively before *t*, such as for a example a decrease (A) or increase (B) of the number of family members in a household. Constraints were included in the structure learning task to prevent that directions of links are back-in-time. Furthermore, the following constraints are defined on the basis of a distinction between external variables (those that are observed when a prediction is to be generated) and target variables (those that are to be predicted). In this case, the target variables are the Occurrence event nodes and the other nodes are the external variables. All external variables can only have a link with target variables Occurrence event within an Event or across an Event. Links between the external variables are not allowed. The external variables age and gender can only have *outgoing* links to the target variables. This is done to prevent the creation of large underlying conditional probability tables, while the direction of links has no implications for predictions. Links that represent logical relationships are enforced.

The learned Bayesian belief network included all seven structural life course events: change in residential location (in short, housing), change in household composition (household), change in work location (work), change in study location (study), change in car possession and availability (availability), change in availability of public transport pass (PT pass), and change in household income (household income). Figure 1 portrays the network for the lifcycle events, which includes 42 nodes and 75 links (42 of which were predefined).

The 33 learned links were distributed across these five categories: 15 links were learned within an event, 5 links across events, 8 links between the event nodes, 4 links with the variable age, and 1 link was learned between the personal characteristics.

Figure 2 portrays the second network, which examines if and how life trajectory events impact transport mode choice. Transport mode choice was defined as the most frequently used mode across the activities of the person. The network has 43 nodes and 79 learned links: 75 links were already learned in the first network and 4 extra links were learned with the node mode choice.

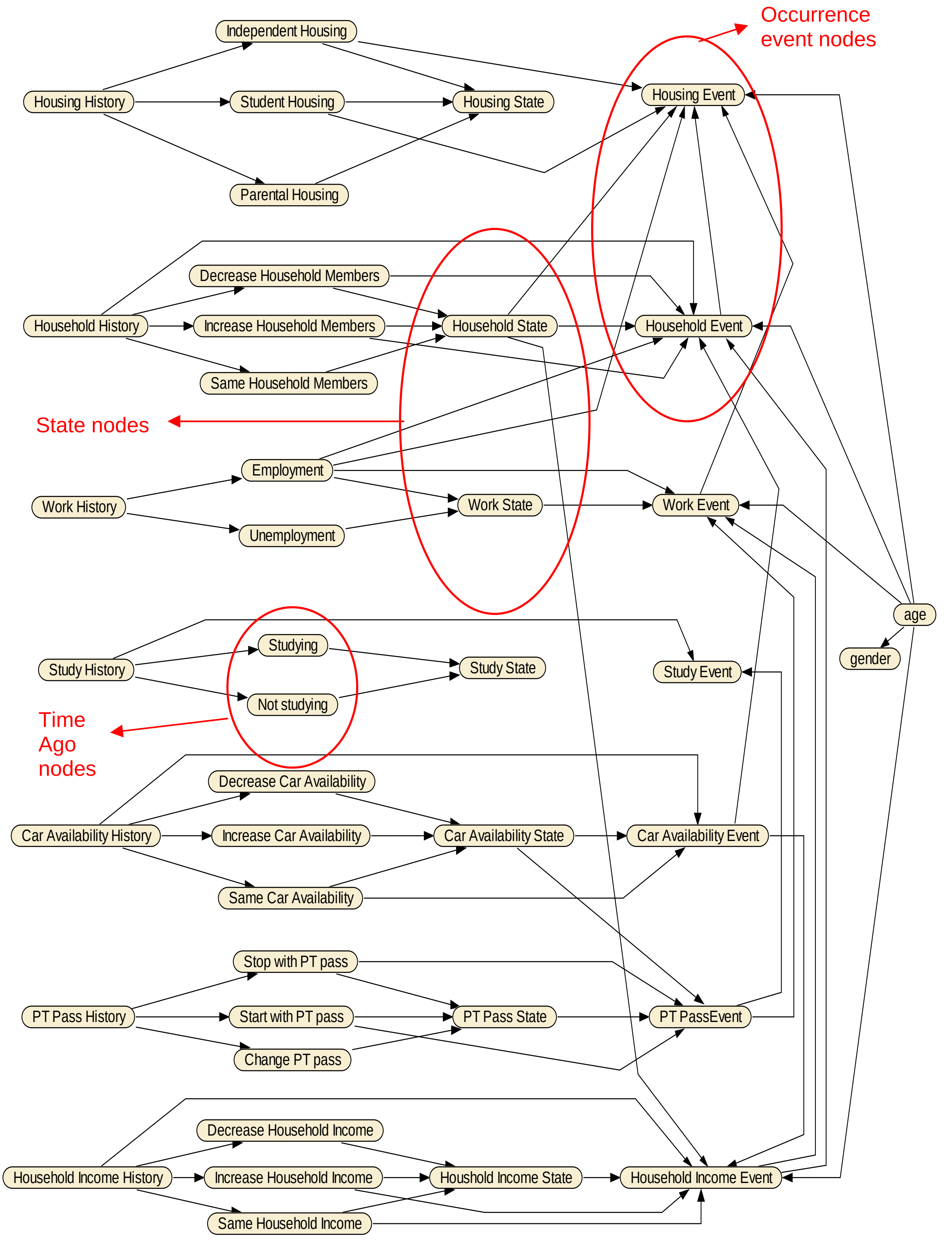

Figure 1 *– Learned life trajectory network*

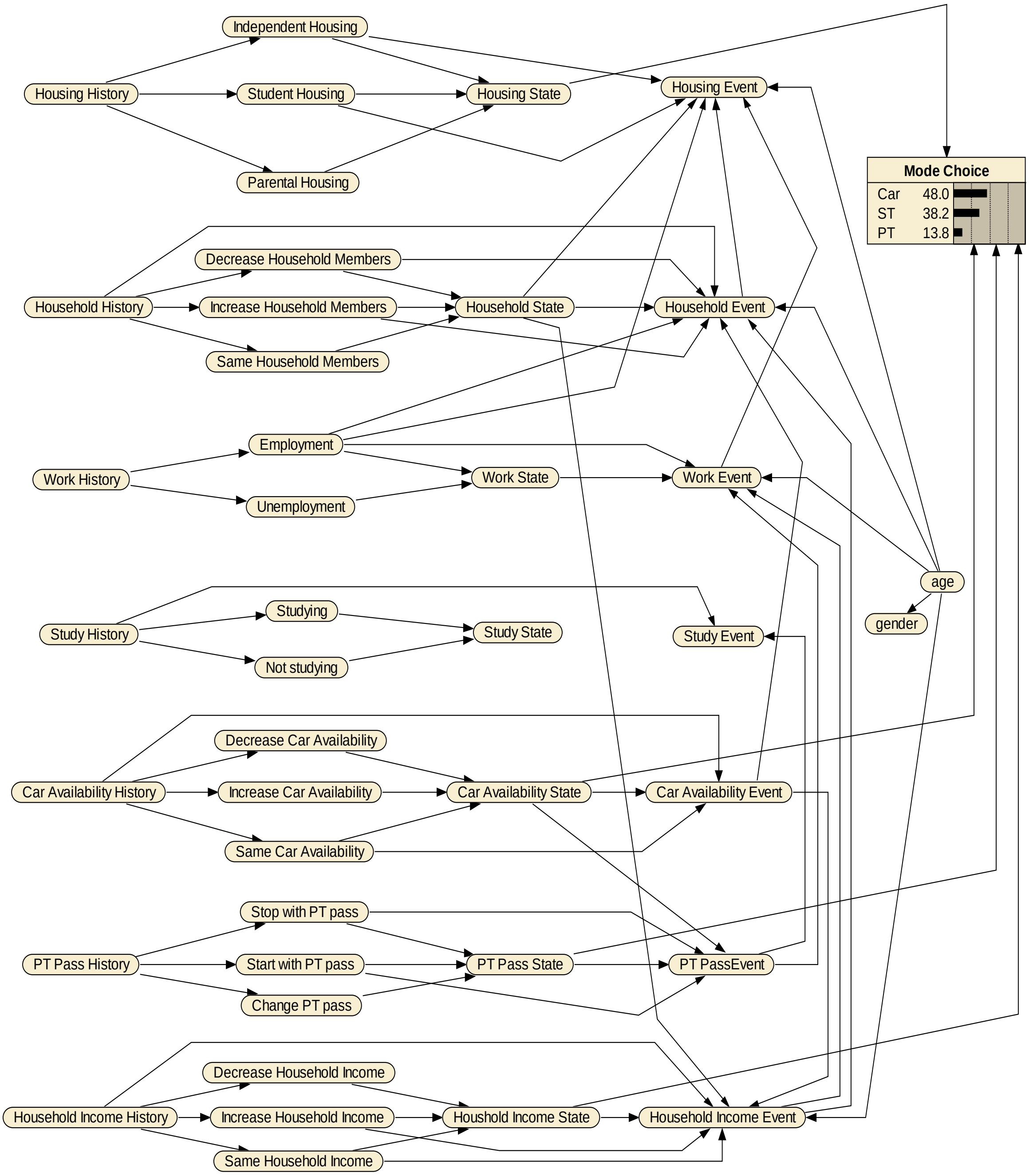

Figure 2 *– Learned mode choice network*

VALIDATION OF THE LIFE TRAJECTORY NETWORK

The learned networks provide a general representation of the data. The graphs however provide little information about the extent to which the predictions of the network are consistent with the observed life trajectories and current mode choice. To examine the validity of the networks, the following comparisons between predicted and observed life trajectories were made: (i) number of occurrences (in short 'count'), (ii) interval times between occurrences of events (in short ‘interval time’), (iii) simultaneous occurrences of events (in short ‘synchronic events), and (iv) sequence of occurrences of events (in short ‘sequence’).

The occurrence of an event in year *t* depends on the predicted probabilities for that event given all variables/nodes in the network. Monte Carlo simulation is used to sample a specific state for the occurrence event node. Only one state of an event can be chosen in one year. A year is chosen to represent time interval; within a year there is no order between occurrences.

Because the process is probabilistic, different simulation runs lead to different results. For validation purposes, five simulations were run. The input for every simulation run is the same: a set of 700 cases. Each case corresponds with one respondent in the sample. All life trajectories are simulated until the year 2004.

Goodness-of-fit

To assess the goodness-of-fit of the learned network model regarding the prediction of events, the log-likelihood is calculated for each event type separately, based on a prediction sequence. The sequence in which event types are predicted is predefined based on their mutual influences as revealed by the network structure. The sequence starts with an event with no incoming links from other event nodes and continues with events which only have links from the event nodes which precede in the sequence. The observed outcome of an event is taken as given in the prediction of the next event. Log likelihood values were calculated for three models: null-model, overall model and prediction model. In the null-model uniform distributions were chosen, implying that every option has the same chance. In the overall model the probabilities were distributed according to the overall probability distribution in the data set. The probabilities in the prediction model were based on the prediction based on the network. The log likelihood was calculated for every event and for every model. The calculated log likelihoods are listed in Table 1

Table 1 *–Log likelihood values mode choice network*

	Event	Null-model	Overall	Prediction	First	Second
7	**Housing**	-10602.40	-3193.79	-1819.90	0.8284	0.4302
6	**Household**	-10602.40	-3767.82	-1883.04	0.6843	0.5002
5	**Work**	-8402.19	-3422.19	-2652.44	0.9674	0.2249
3	**Study**	-8402.19	-417.56	-273.94	0.8224	0.3440
2	**Car Availability**	-10602.40	-2714.49	-2162.44	0.7960	0.2034
1	**PT pass**	-10602.40	-1645.47	-1183.69	0.8884	0.2806
4	**Household Income**	-10602.40	-3184.56	-1749.95	0.8349	0.4505

Two Rho-Squares were calculated, the first one is the ratio between the log likelihood of the prediction and null-model, and the second one is the ratio between the log likelihood of the prediction and overall model. The values of the first Rho-Squares in Table 1 are relatively high, all above 0.68. This indicates that the Bayesian network model is a strong improvement of the null-model. The values of the second Rho-Square are lower, but all values are above 0.20. This indicates, according to generally accepted norms, that the models perform well.

Counts

The first validation test is whether the total number of years with an occurrence is successfully reproduced by the network. The life trajectories of the 700 individuals in the sample produce 7648 years (i.e. person-years) in total. In every year an occurrence could happen or not. The frequency is the sum of all person-years when an occurrence happened across the (sub)events. The observed frequencies are compared with the frequencies of the simulations. Two classification levels are distinguished for the frequencies:

1. No occurrences/occurrences on the level of subevents

2. No occurrences/occurrences on the level of main events

All years are taken into account in both levels, including the person-years where no occurrence happened. At the first level all available information is taken into account. That is, the level of detail of the subevent is also included. Chi-Square was calculated to test whether the model is capable of reproducing the observations. Table 2 provides the results for the first level: no occurrences/occurrences (subevent level). The first column lists the event with all states (no occurrence and all subevents). The second column shows the total number of counts in the observed life trajectories and the third column presents the percentage of the total count within each event. The total counts for the predicted life trajectory is shown in the fourth column and the percentage of the total count within an event is listed in the fifth column. The Chi-Square and its p-value are listed in the last two columns.

A p-value < 0.05 (if alpha is 5% is applied) means that the null-hypothesis is rejected, and hence that the observed and simulation frequencies differ significantly. The number of person-years with a decrease and increase in the number of household members is underestimated in the simulation. For all other events the simulation produces the same percentage of person-years with and without an occurrence. The occurrence is defined here as subevents, this means that the number of person-years with a specific subevent is predicted.

Table 2 *–Results count level 1 (subevent level)*

Housing	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6708	0.877	33731	0.882		
independent housing	850	0.111	3980	0.104		
student housing	71	0.009	420	0.011		
parental housing	19	0.002	109	0.003		
Total	7648	1.000	38240	1.000	**52.263**	0.156
Household	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6654	0.87	33691	0.881		
Decrease	310	0.041	1334	0.035		
Increase	627	0.082	2920	0.076		
Same	57	0.007	295	0.008		
Total	7648	1.000	38240	1.000	**92.178**	0.027
Work	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6520	0.853	32501	0.85		
employed	1072	0.14	5445	0.142		
unemployed	56	0.007	294	0.008		
Total	7648	1.000	38240	1.000	**0.384**	0.944
Study	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	7583	0.992	37931	0.992		
studying	24	0.003	115	0.003		
not studying	41	0.005	194	0.005		
Total	7648	1.000	38240	1.000	**0.1401**	0.987
Car availability	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	7057	0.923	35181	0.92		
Decrease	150	0.02	701	0.018		
Increase	259	0.034	1377	0.036		
Same	182	0.024	981	0.026		
Total	7648	1.000	38240	1.000	**23.044**	0.512
PT pass	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	7336	0.959	36458	0.953		
Stop	100	0.013	602	0.016		
Start	116	0.015	626	0.016		
Change	96	0.013	554	0.014		
Total	7648	1.000	38240	1.000	**54.444**	0.142
Household Income	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6847	0.895	34215	0.895		
Decrease	43	0.006	235	0.006		
Increase	215	0.028	1028	0.027		
Same	543	0.071	2762	0.072		
Total	7648	1.000	38240	1.000	**0.7788**	0.8545

Fout! Verwijzingsbron niet gevonden. Table 3 reports the results for the second level: no occurrences/ occurrences on the main event level. On this level, 2 of the events, namely household and PT pass, have a p-value below 0.05. In case of the household event, the person-years with occurrences are underpredicted in the simulation and for the PT pass event the person-years with occurrences are overpredicted. In general, the life trajectory model reproduced the number of person-years with and without occurrences in the life trajectories quite well.

Table 3 *–Results count level 2 (main event level)*

Housing	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6708	0.877	33731	0.882		
occurrence	940	0.123	4509	0.118		
Total	7648	1.000	38240	1.000	**15.194**	0.218
Household	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6654	0.87	33691	0.881		
occurrence	994	0.13	4549	0.119		
Total	7648	1.000	38240	1.000	**72.738**	0.007
Work	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6520	0.853	32501	0.85		
occurrence	1128	0.147	5739	0.15		
Total	7648	1.000	38240	1.000	**0.3357**	0.562
Study	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	7583	0.992	37931	0.992		
occurrence	65	0.008	309	0.008		
Total	7648	1.000	38240	1.000	**0.138**	0.710
Car availability	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	7057	0.923	35181	0.92		
occurrence	591	0.077	3059	0.08		
Total	7648	1.000	38240	1.000	**0.6439**	0.422
PT pass	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	7336	0.959	36458	0.953		
occurrence	312	0.041	1782	0.047		
Total	7648	1.000	38240	1.000	**49.322**	0.026
Household Income	**observed**	**%**	**predicted**	**%**	**Chi Square**	**p-value**
no occurrence	6847	0.895	34215	0.895		
occurrence	801	0.105	4025	0.105		
Total	7648	1.000	38240	1.000	**0.0185**	0.892

Interval times

This second validation test analyses the interval times between occurrences within one event and between occurrences of two different events. The timing of occurrences is analysed given two interval types. Mean values are calculated for every interval type. An independent-samples *t*-test is used to test whether the means of the two independent random samples (observed and predicted sample) are statistically different. It is important to note that, in this analysis, there is no distinction between the occurrences on the level of subevents, implying that all subevents as mentioned before are listed as occurrences without further qualification.

Three intervals are distinguished here: start interval, sequence interval and end interval. While only one interval (i.e. sequence interval) was used in the analysis between the observed and predicted life trajectories. The interval from the start to the first occurrence is named start interval. In case there was an occurrence in the first year of the life trajectory there is no start interval. The same holds for the end interval, which is the interval between the last and most recent occurrence and the end. These two intervals are not taken into account in this analysis. The start and end moment of the life trajectory are completely random. Both are determined in the survey, the start interval can be defined by the occurrence of another (related) event, by routing of the survey, the end interval is established by the moment of the data collection. The first occurrence in the life trajectory is not predicted given the previous occurrence, because this occurrence is not known. For these reasons, the start and end interval are not taken into account here.

The interval between two occurrences of the same event is called sequence interval. If there are no occurrences in a person's life trajectory no sequence intervals are registered.

Only the mean sequence intervals and the mean interval between two events are analysed. Table 4 lists mean and number of the intervals of the observed life trajectories and Table 5 lists this information for the predicted life trajectories. The sequence intervals are in the diagonal of the table and the intervals between any two different types of events are in the other cells of the table. The upper part of the tables contains the mean values and the lower part lists the total number of intervals. The totals of Table 5 represent a sum across five simulations.

The null hypothesis (H_0) is that the model predicted the observed means of intervals well ($\mu_1 = \mu_2$). The independent sample *t*-test tests whether the predicted and observed means are equal. Table 6 lists the results of the *t*-test.

Table 4 *–Interval times (observed data)*

Means	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing	4.5	4.4	3.4	3.2	5.4	3.5	3.8
HH	5.2	4.5	4.3	3.8	5.7	4.2	5.1
Work	5.0	5.4	3.9	2.8	6.2	4.1	5.2
Study	1.6	2.1	1.7	2.3	1.7	1.9	1.9
Car	5.6	5.9	4.8	3.3	5.2	4.0	5.6
PT pass	5.5	4.8	4.1	3.5	6.7	3.8	5.1
Income	4.6	5.2	4.3	3.0	5.6	3.8	4.8
intervals	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing	531	466	549	45	337	158	453
HH	615	629	621	29	445	178	524
Work	684	596	740	44	424	188	539
Study	20	14	12	7	14	10	14
Car	329	309	350	27	279	126	322
PT pass	177	149	185	31	112	120	165
Income	435	400	434	34	311	148	457

Table 5 *–Interval times (predicted data)*

Means	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing	4.7	4.5	3.7	2.4	5.7	4.7	4.6
HH	5.4	4.7	4.5	2.8	6.6	5.7	5.4
Work	5.4	5.2	4.4	2.2	6.4	5.1	5.3
Study	2.2	2.1	2.2	2.2	2.8	2.4	2.4
Car	5.9	5.6	5.0	1.9	5.5	5.6	5.9
PT	5.6	5.3	4.3	2.2	6.3	4.4	5.3
Income	5.0	4.9	4.4	2.6	6.1	4.9	4.9
intervals	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing	2510	2117	2801	105	1816	727	2192
HH	2728	2817	3043	72	2068	833	2477
Work	3206	2949	3792	70	2363	1077	2889
Study	104	73	111	28	69	42	88
Car	1569	1513	1710	37	1381	558	1399
PT	1011	889	1143	53	715	827	929
Income	1958	1824	2083	69	1488	597	2175

***Table 6** –Results* t-*test for all interval times*

	Housing	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing	0.18	0.45	0.11	**0.03**	0.36	**0.00**	**0.00**
HH	0.46	0.28	0.22	0.08	**0.002**	**0.00**	0.32
Work	0.10	0.28	**0.006**	0.10	0.47	**0.001**	0.65
Study	0.07	0.91	0.21	0.91	**0.003**	0.29	0.29
Car	0.45	0.36	0.37	**0.006**	0.45	**0.00**	0.33
PT pass	0.89	0.26	0.53	**0.015**	0.55	0.11	0.65
Income	0.17	0.23	0.55	0.28	0.13	**0.001**	0.83

All significance values below 0.05 are bold in Table 6. In these cases the null hypothesis is rejected (alpha of 5% is applied), which means that the difference between the observed mean intervals and the predicted mean intervals is significant. The difference between the observed mean intervals and the predicted mean intervals is significant for 12 out of 49 intervals. In general, three interval times are underestimated and the other nine intervals are overestimated.

In the sequence interval group there is one overestimation. The observed interval for the Work event is 3.9 years while the predicted interval is 4.4 years. A total of 11 out of 42 intervals in the last group (i.e. intervals between two events) are significantly different. The following intervals are underestimated: housing and study, car availability and PT pass, and PT pass and study. The model overestimates the interval between housing and PT pass, housing and household income, household and car availability, household and PT pass, study and car availability, car availability and study, and household income and PT pass.

There are more overestimated intervals than underestimated intervals. This indicates that the predicted time interval between two events on average is longer than reported in the survey. In general, the model predicts more or less the same interval times for the other events. The most difficult interval to predict is the interval between an event and the PT pass event. The model needs further improvement especially for this event.

Synchronic events

A time interval does not exist when two events occur in the same year. The time order within one year is no longer traceable due to recoding of the data according to the chosen time unit of one year. If occurrences were reported in the same year in the Internet-based survey, it is important that the network can reproduce this synchronism.

Table 7 shows the number of synchronic events. Table 8 lists the number of observed occurrences for each event with no event occurrence as well as occurrences with one or more other event occurrences. The total sum here is the sum of the previous rows; this exceeds the total number of occurrences reported. The last row in the table indicates the number of occurrences reported in the Internet-based survey (see table 3). The probability of synchronic events depends on the total number of events. For example, when a housing event happened, the probability of a household occurrence in the same year is the ratio between synchronic events (housing and household) and the total number of occurrences (housing). In this example this is 288 / 940 = 0.31. Table 9 lists all probabilities. In Table 10 – Table 12 the same information is listed for the predicted events: synchronic events, the number of synchronic events, probabilities of synchronic events.

Table 7 *–Synchronic events frequencies (observed data)*

Observed	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing		288	271	32	178	87	240
HH	288		213	12	207	60	276
Work	271	213		27	160	145	359
Study	32	12	27		14	25	27
Car	178	207	160	14		66	186
PT pass	87	60	145	25	66		112
Income	240	276	359	27	186	112	

Table 8 *–Synchronic events more than two in one year (observed data)*

Observed	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
no other event	384	426	457	13	196	75	186
1 event	236	272	361	8	155	92	252
2 events	168	165	180	20	120	77	206
3 events	96	81	78	13	74	34	104
4 events	45	40	41	6	37	24	42
5 events	10	9	10	4	8	9	10
6 events	1	1	1	1	1	1	1
Total	**1096**	**1056**	**1175**	**137**	**811**	**495**	**1200**
Occurrences	**940**	**994**	**1128**	**65**	**591**	**312**	**801**

Table 9 *–Synchronic events probabilities (observed data)*

Observed	**Housing**	**HH**	**Work**	**Study**	**Car**	**PT**	**Income**
Housing	0.00	0.31	0.29	0.03	0.19	0.09	0.26
HH	0.29	0.00	0.21	0.01	0.21	0.06	0.28
Work	0.24	0.19	0.00	0.02	0.14	0.13	0.32
Study	0.49	0.18	0.42	0.00	0.22	0.38	0.42
Car	0.30	0.35	0.27	0.02	0.00	0.11	0.31
PT pass	0.28	0.19	0.46	0.08	0.21	0.00	0.36
Income	0.30	0.34	0.45	0.03	0.23	0.14	0.00

Table 10 *–Synchronic events frequencies (predicted data)*

Simulations	Housing	HH	Work	Study	Car	PT	Income
Housing		1021	1134	59	529	252	820
HH	1021		838	50	615	208	966
Work	1134	838		53	668	588	1574
Study	59	50	53		32	85	39
Car	529	615	668	32		151	805
PT	252	208	588	85	151		211
Income	820	966	1574	39	805	211	

Table 11 *–Synchronic events more than two in one year (predicted data)*

Observed	Housing	HH	Work	Study	Car	PT	Income
no other event	2110	2174	2523	118	1373	758	1303
1 event	1404	1435	2012	108	923	677	1475
2 events	649	630	847	51	484	244	875
3 events	276	242	284	22	212	86	302
4 events	65	63	68	8	62	13	66
5 events	5	5	5	2	5	4	4
6 events	0	0	0	0	0	0	0
Total	**4509**	**4549**	**5739**	**309**	**3059**	**1782**	**4025**
Occurrences	**3815**	**3698**	**4855**	**318**	**2800**	**1495**	**4415**

Table 12 *–Synchronic events probabilities (predicted data)*

Observed	Housing	HH	Work	Study	Car	PT	Income
Housing	0.00	0.23	0.25	0.01	0.12	0.06	0.18
HH	0.22	0.00	0.18	0.01	0.14	0.05	0.21
Work	0.20	0.15	0.00	0.01	0.12	0.10	0.27
Study	0.19	0.16	0.17	0.00	0.10	0.28	0.13
Car	0.17	0.20	0.22	0.01	0.00	0.05	0.26
PT pass	0.14	0.12	0.33	0.05	0.08	0.00	0.12
Income	0.20	0.24	0.39	0.01	0.20	0.05	0.00

Table 13 *–Results p-values for the binomial test*

Observed	Housing	HH	Work	Study	Car	PT	Income
Housing		0.000	0.000	0.000	0.000	0.000	0.000
HH	0.000		0.000	**0.270**	0.000	0.000	0.000
Work		0.000		0.000	0.000	0.000	0.000
Study	0.000	**0.226**	0.000		0.000	0.000	0.000
Car	0.000	0.000	0.000	0.000		0.000	0.000
PT pass	0.000	0.000	0.000	0.000	0.000		0.000
Income	0.000	0.000	0.000	0.000	0.000	0.000	

Table 13 represents the results of binomial test. A p-value smaller than 0.05 indicates a significant difference between the observed and predicted probability of synchronic events. In almost all cases there is a significant difference. The probability of the synchronic events Household and Study is an exception. The values 0.270 and 0.226 indicate that there is no significant difference. This means the model is successful in predicting these synchronic events. In general the model is less successful in predicting correctly the observed synchronic events.

Sequence (SAM)

Occurrences in a life trajectory take place in a certain order or sequence. In order to compare the sequence of the observed life trajectory with the life trajectory of the simulations, the Sequence Alignment Method (SAM) was used. Three different sequences were constructed and analysed. The first sequence option (labelled as OccurrenceOnly) is constructed with no reference to a year. This means that only the occurrences are listed in the sequence. There is no distinction in subevents in the sequence. The next sequence (labelled as SubstatesOnly) is a more detailed description of the first sequence in that the subevents are listed. There is still no reference to the year of occurrence, just like in the OccurrenceOnly sequence. The last sequence (labelled as Life Trajectory) is a multidimensional string where each dimension corresponds to an event type and a string indicates yes/no occurrence for each year in a row.

Sequence alignment methods calculate the similarity between two sequences in terms of the number of operations that is required to equalize the two sequences. The operations involve adding, deleting, and substitution. In principle, different weights can be attached to these operations. Sequence alignment methods were originally developed for uni-dimensional sequences (Kruskal, 1983; States and Boguski, 1991). Joh, Arentze and Timmermans (2001) further expanded these methods to the case of multidimensional sequences. The Life Trajectory sequence can be viewed as a multidimensional profile. In this case, two different sequence alignment methods were applied: UDSAM (the sum of uni-dimensional sequence alignments) and DPSAM (Dynamic programming-based multidimensional sequence alignment method (Joh *et al*., 2001)). The UDSAM measure calculates the alignment costs of each uni-dimensional sequence and the sum of these measures across all dimensions is taken as a measure of distance or dissimilarity between two multidimensional patterns. The DPSAM measure compares sequences on multiple dimensions simultaneously, taking dependencies between the dimensions into account.

In our analyses, a weight of one was used for additions and deletions, while swapping received a weight of two (it can be seen as a combined deletion and addition operation). Thus, the SAM measure for each sequence alignment is the weighted total number of additions, deletions and substitution operations required to make the two sequences identical. A higher number implies that the two sequences are less identical.

The value of SAM (costs to align the sequences) will increase when the level of detail increases. It is difficult to analyse the SAM value if there is no comparison or range from the minimum to maximum SAM value. For this reason, a minimum and maximum were calculated for each observation and the mean of all SAM values was compared to this range.

Given the three sequences, four analyses were executed. The UDSAM measure is used for the OccurrenceOnly sequence and the SubstatesOnly sequence. Both sequences are uni-dimensional. The Life Trajectory sequence is analysed two times: first as an entity with the DPSAM method (Dynamic programming-based multidimensional sequence alignment method) and second for each event separately with UDSAM method (uni-dimensional). The Life Trajectory sequence is multi-dimensional given the seven strings which represent the different time lines in the life trajectory. The range of the SAM value is defined as follows. The minimum, obviously, is always equal zero. The definition of the maximum depends on the measure considered. For the OccurenceOnly and SubstatesOnly, the maximum is defined as: length sequence x 2, whereas for the Life Trajectory it is defined as: SAMmax = length sequence x 2 x 7. The multiplication with 2 is for substituting a value in the sequence and the multiplication with 7 is the number of events. The length of each string in the Life Trajectory sequences is the same for the observed and predicted sequences. If a person reported a life trajectory of twelve years the predicted life trajectory also consists of twelve years. In OccurenceOnly and SubstatesOnly sequences, the length of the string is the maximum length of observed and predicted string. The SAM value is indicated with a percentage, calculated as: percentage SAM value = (SAM value - SAM $\text{value}_{\text{min}}$) / SAM $\text{value}_{\text{range}}$

The lower the percentage the closer the SAM value is to the minimum SAM value, thus the better the performance. In total 3495 cases were analysed. In some cases, the length of the observed or predicted sequence is zero. This means that no occurrences were observed or predicted for that person. These cases, in total 723, were not included in the analysis. In total, 2772 cases were analysed.

Table 14 *–Results OccurrenceOnly sequence (UDSAM)*

	Minimum	**Maximum**	**Mean**	**Std. Deviation**
Length sequence	1	31	8.54	5.83
Minimum SAM	0	28	3.73	3.73
Maximum SAM	2	62	17.08	11.66
SAM range	2	60	13.35	10.14
SAM value	0	31	8.48	5.74
SAM percentage	0	1	0.35	0.26

***Table 15** –Results SubstateOnly sequence (UDSAM)*

	Minimum	Maximum	Mean	Std. Deviation
Length sequence	1	31	8.54	5.83
Minimum SAM	0	28	3.73	3.73
Maximum SAM	2	62	17.08	11.66
SAM range	2	60	13.35	10.14
SAM value	0	37	9.79	6.69
SAM percentage	0	1	0.45	0.28

The results of the analysis of the OccurrenceOnly sequence option are listed in Table 14, while Table 15 shows the results for the SubstatesOnly sequence option. Of course, the values for the length of the sequences, the minimum and maximum SAM value and the SAM range are the same for the OccurrenceOnly and SubstatesOnly sequence option. Those values are listed in the first four rows of the tables.

The percentage of the analysis of OccurrenceOnly and SubstatesOnly sequence are both below 0.5. The mean percentage of the SubstatesOnly sequence option (0.45 in Table 15) is higher than the mean percentage of the OccurrenceOnly sequence option (0.35 in Table 14). This can be explained by the difference in level of detail between the two sequence options. More detail in the sequence means a higher probability of differences between the sequences, which leads to a higher SAM (and SAM percentage) value.

SAM values for the Life Trajectory sequence option were calculated for 3495 cases. The total length among the 3495 cases varied from 1 year to 40 years. Table 16 report the results of the Life Trajectory sequence.

The mean value of the total SAM value based on the uni-dimensional (UDSAM) method is higher (15.91) than the mean value based on the multidimensional method (10.35). In general, the MDSAM value is always below the sum of UDSAM or it can exactly match the sum. The mean value of the SAM percentage (DPSAM) is close to zero. The last column in **Fout! Verwijzingsbron niet gevonden.**16 lists the number of cases with a SAM value of zero. The results indicate that the Study event and the PT pass event were simulated most closely by the Bayesian belief network.

Highest SAM measures were obtained for the work event and the household event, suggesting that the Bayesian network was relatively less successful in correctly simulating the timing of occurrences and/or sequence of these event types.

Based on the SAM measures, it is difficult to tell whether mis-predictions are caused by (1) predicting the wrong subevent or (2) predicting a wrong timing of the subevent. Overall, however, the results of the SAM analyses demonstrate that the learned Bayesian belief network predicts the sequence of the occurrences in the life trajectories relatively well.

***Table 16** –Results Life Ttrajectory sequence (UDSAM)*

DPSAM analyses	**Minimum**	**Maximum**	**Mean**	**Std.**	**Zero SAM**
Length sequence	1	40	10.93	10.37	
Minimum SAM	0	0	0.00	0.00	
Maximum SAM	14	560	152.98	145.12	
DPSAM analyses					
SAM value	0	49	10.35	9.04	
SAM percentage	0	0.14	0.08	0.04	
UDSAM analyses					
SAM value Housing	0	20	2.75	2.83	1179
SAM value Household	0	18	3.00	3.32	1393
SAM value Work	0	20	3.24	3.59	1369
SAM value Study	0	6	0.32	0.83	2984
SAM value Car	0	26	2.36	2.73	1353
SAM value PT pass	0	16	1.54	2.33	2033
SAM value Income	0	20	2.70	3.11	1422
SAM value total	0	66	15.91	13.68	

VALIDATION OF THE MODE-CHOICE NETWORK

Mode choice was only registered in the year of the retrospective Internet-based survey (2004). This means that the observed data for mode choice is limited to one year. In this section, we test the validity of the second network for predicting mode choice. First, the log likelihood will be discussed for three models: null-model, overall model and prediction model. Next, the results of a simulation will be discussed as an illustration.

Goodness-of-fit

One way of assessing the validity of the learned network is to use the values for all variables/nodes in the network in the year 2004, except mode choice, and predict the posterior probabilities of a particular transport mode choice in the year 2004, given this hard evidence. In this case, the hard evidence corresponds with the situation in the corresponding year.

The network was used to calculate the posterior probabilities for every case given the hard evidence. To test the performance of this mode choice model, the log likelihood value for three models (null-model, overall model and prediction model) was calculated. Table 17 lists the values for the different models. In the overall model, the probabilities were distributed according to the overall probability distribution. Two Rho-Squares were calculated, the first one is the ratio between the log likelihood of the prediction model and null model, and the second one is the ratio between the log likelihood of the prediction model and overall model. The values of both Rho-Squares are reported in Table 17. Both Rho-Square values of the model are above 0.36, which indicate that the model performs relatively well.

Table 17 *–Log-likelihood values mode choice network*

	Null model	Overall	Prediction	First Rho-Square	Second Rho-Square
Mode Choice	-769.03	-697.57	-439.84	0.4281	0.3695

Mode choice in 2004

Monte Carlo sampling is used to select an individual's mode choice based on the predicted posterior probabilities. The predicted modal split can be calculated given the individuals' mode choices. The observed modal split in 2004 is compared with predicted modal split to show how the model can be used for simulation. The results are shown in Table 18 and 19.

Only one Monte Carlo simulation was run. In 64 percent of the cases the predicted individual's mode choice corresponds with the observed individual's mode choice. Table 19 reports a relatively small overprediction of public transport at the expense of the other two mode choices car and slow transport.

Table 18 *–Mode choice in 2004 (observed and predicted)*

		observed mode			
		Car	**PT**	**ST**	**Total**
predicted mode	**Car**	199	69	23	**291**
	PT	70	220	37	**327**
	ST	23	30	29	**82**
	Total	**292**	**319**	**89**	**700**

Table 19 *–Modal split in 2004 (observed and predicted)*

Mode	observed	%	predicted	%	Difference	prediction
Car	292	0.417	291	0.416	-0.001	Under prediction
PT	319	0.456	327	0.467	0.011	Over prediction
ST	89	0.127	82	0.117	-0.010	Under prediction
Total	700	1.000	700	1.000		

CONCLUSIONS AND DISCUSSION

To contribute to the literature on modeling the impact of long-run life course events on travel decisions, this paper has reported the main results of several validity tests of two related Bayesian belief networks that were learned from retrospective data to simulate life trajectories and predict these effects. The log likelihood values for the prediction of the occurrence of events and mode choice suggest that the learned networks perform relatively well. Results indicate that the life trajectory model reproduced the number of occurrences in the life trajectories quite well. The model predicted more or less the same interval times for the events, except for the PT pass event. The results of the validation tests showed that the model was less successful in predicting correctly the observed synchronic events, while the results of the Sequence Alignment Method demonstrated that the Bayesian belief network more or less reproduced the sequence of events as registered in the observed life trajectory.

For the mode choice network, a further illustration involved examining whether modal split was predicted correctly in the simulation. Results of this illustration indicated a small overprediction of public transport and underprediction of car and slow transport. This suggests that the mode choice network can reproduce more or less observed mode choice.

Overall these findings suggest that a representation of life trajectories in terms of Bayesian belief network may be a valuable approach. Our study presented and illustrated a methodology, in the form of a set of measures, to test the validity of a Bayesian belief network model for predicting life trajectories of individuals. The tests reported provide some evidence that the specific model considered here performs satisfactory. We emphasize, however, that for a more definite test validation should be based on a hold-out set which, given data limitations (sample size), was not performed here.

REFERENCES

Andersen, S.K., K. G. Olesen, F. V. Jensen & F. Jensen (1989). Hugin – a shell for building Bayesian belief universes for expert systems. In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, pages 1080-1085, Detroit, Michigan, Aug. 20-25.

Beige, S. (2008). Long-term and mid-term mobility decisions during the life course. Dissertation, Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, Zürich.

Joh, C.H., Arentze, T.A. and Timmermans, H.J.P. (2001). Pattern recognition in complex activity-travel patterns: A comparison of Euclidean distance, signal processing theoretical, and multidimensional sequence alignment methods, *Transportation Research Record*, 1752, 16-22

Kruskal, J.B. (1983). An overview of sequence comparison. In: Sankoff, D., Kruskal, J.B. (Eds.), Time Warps, StringEdits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison-Wesley, London, pp. 1–44.

States, D.J., and Boguski, M.S. (1991). Similarity and homology. In: Gribskov, M., Devereux, J. (Eds.), Sequence Analysis Primer. Stockton Press, New York, pp. 89–157.

Verhoeven, M., Arentze, T.A., Timmermans, H.J.P. and van der Waerden, P.J.H.J. (2005). Modelling the Impact of Key Events on Long-Term Transport Mode Choice Decisions: A Decision Network Approach Using Event History Data. In *Transportation Research Record*: Journal of the Transportation Research Board, No. 1926, TRB, National Research Council, Washington, D.C., 2005, pp. 106-114.

Verhoeven, M., Arentze, T.A., Timmermans, H.J.P. and van der Waerden, P.J.H.J. (2006). Modelling the Influence of Structural Lifecycle Events on Transport Mode Choice Decisions Using a Structure Learning Algorithm. *Proceedings 11th IATBR Conference* (CD-ROM), Kyoto, Japan.

Verhoeven, M., Arentze, T.A., Timmermans, H.J.P. and van der Waerden, P.J.H.J. (2007). Examining temporal effects of lifecycle events on transport mode choice decision. The International Journal of Urban Sciences. vol. 11. 1-14.