# SEMI-AUTOMATIC IMPUTATION OF LONG-TERM ACTIVITY-TRAVEL DIARIES USING GPS TRACES: PERSONAL VERSUS AGGREGATE HISTORIES

*Anastasia Moiseeva*
*Faculty of Architecture, Building and Planning - Eindhoven University of Technology - P.O. Box 513 - 5600MB Eindhoven - The Netherlands – A.Moiseeva@tue.nl*

*Harry J. P. 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*

*Joran Jessurun*
*Faculty of Architecture, Building and Planning - Eindhoven University of Technology - P.O. Box 513 - 5600MB Eindhoven - The Netherlands - A.J.Jessurun@tue.nl*

## ABSTRACT

The new generation of dynamic models of activity-travel demand requires multi-day or multi-week activity-travel data. A combination of modern GPS technology and a prompted recall instrument may be a powerful tool to reduce respondent burden and collect such multi-week data of travel behavior. The authors have developed such a system, called TraceAnnotator which was designed for automatic imputation of various facets of activity-travel patterns form GPS tracers. The core of the system developed is a Bayesian belief network that classifies the outcome variables of interest, using a network of input variables. This means that the interpretation of the GPS traces of any new respondent is based on the aggregate conditional probability tables that the system learned on the basis of the previously processes respondents/cases. However, in case of multi-day data collection, the history of every respondent is also collected. This implies that learning can be based on the continuously updated conditional probability tables, aggregated across respondents, or on the personal histories of respondents or on a combination of both.  This paper will discuss the results of these alternative approaches to impute transport modes and activity types for multi-week activity-travel diary data using GPS technology.

*Keywords: multi-day GPS tracers, personal data, aggregate data, learning algorithm*

# INTRODUCTION

The last decade has witnessed a rapidly increasing number of academic studies applying modern technology, such as GPS, to collect activity-travel data (see Timmermans et al., 2009, for a recent overview). GPS and similar technology allows tracing respondents. Data on timing and duration, route choice and destination choice is relatively easily obtained, although data collection is not necessarily without errors. In any case, the potential of this new technology has been realized and the first larger scale, non-academic data collections are in the field.

Moreover, several other travel characteristics, such as number of trips, activity stops, duration and transport mode can be derived from GPS traces. However such on transport mode choice and activities GPS tracers do not provide per se. In urban and transportation planning there have been attempts to derive these kinds of data from the characteristics of the GPS traces, in combination with land use data. Most researchers have used ad hoc rules and researcher-driven decisions to extract additional travel features from GPS traces (Marca et al, 2002; Axhausen et al, 2004; Chung and Shalaby, 2005; Stopher and Collins, 2005; Doherty et al, 2006; Li and Shalaby, 2008; Bohte and Maat, 2008). Rudloff and Ray (Rudloff and Ray, 2010) compared mode detection algorithms based on decision trees, logistic regression, multilayer perceptrons and support vector machines as classification methods.

The authors have developed a computer environment, called *TraceAnnotator*, that allows such semi-automatic imputation of data about activity-travel patterns, based on GPS traces (Moiseeva et al., 2010). It differs from previous work in transportation research in that the imputation is not based on ad hoc rules, but rather on Bayesian belief networks. The system can with a high accuracy impute facets of travel-patterns. In addition, the system is able to learn and to improve the classification by processing corrections made by respondents in a prompted recall survey and automatically updating the conditional probability tables of the Bayesian belief Network. The learning algorithm is crucial when respondents are required to provide GPS tracers over a longer time period, such as several weeks or months. If the system can efficiently learn over time,  respondent's burden will reduce over time as they need to make less changes. It is expected that corrected and confirmed data from the prompted recall instrument will asymptotically increase the probability of a correct imputation of transport modes and activities with an increasing amount of data.

The specific goal of this paper to test how the imputation of different facets of activity-travel patterns improves over time, in terms of speed and accuracy, if learning process over a long time based on tracers of the same respondent or on aggregated tracers history across respondents.

This paper is organized as follows. First, we will summarize the imputation system TraceAnnotator. Then, we will describe the data and the results of the analysis. Finally, we will draw some conclusions.

# TRACE ANNOTATOR

The general purpose of 'TraceAnnotator' is to semi-automatically process multi-day or multi-week GPS traces. To that effect, a distinction is made between (i) imputation of transportation modes and activity episodes, where data imputation is established by using a Bayesian belief network, and (ii) imputation of activity type, where GPS data are fused with GIS land use data and personalised land use data. TraceAnnotator has been developed using Java and it uses the following technologies: Spring for configuration using xml files (http://www.springsource.org/); GeoTools for the GIS based components (http://geotools.codehaus.org/), and Netica software for the Bayesian Network component used in the implementation of the ClassifierFilter (http://www.norsys.com/).

## Filters

The system has been designed such that it can process arbitrarily large datasets fast, can handle GIS calculations, can be used without programming, and can be easily extended. A filter is the main class of TraceAnnotator. A filter processes a trace of samples, whereas a sample is one measurement from a GPS tracer. Sample contains attributes such as date, time, latitude, longitude and etc. Most filters will manipulate each sample. For example, it can add new attributes or change existing attribute values, but a filter could also write derived data into an external file. Multiple filters can be chained together and each filter can make some changes to the sample or do some data processing and then send the sample to the next filter. By using these filters as building blocks more complex processing can be done without having to program new filters.

The different transport modes and activity episodes cannot be distinguished without additional information such as average acceleration, maximum acceleration, maximum speed and other factors, including errors in the GPS device itself. Again, filters can be used to provide the chain of necessary calculations in order to derive additional information.

## Bayesian belief network

In order to identify transport modes and activity stops from GPS data stream specific rules and several variables were derived positioned on the real observations of the GPS data. Based on these variables, a Bayesian network is used to impute the various facets of the travel pattern.

This network represents the multiple relationships between different spatial, temporal and other factors, including errors in the technology itself and the facets of activity-travel patterns that we wish to impute from the GPS traces. A total of 7 different types of transport modes are considered: walking (by foot), running, bike, motorbike, car, bus and train. It is straightforward to add other modes such as tram or metro if it is necessary. Two types of the activity stops are defined: activity at location (when a person conducts an activity at a certain location) and insidebuilding (when a person conducts an activity inside a building). Initial conditional probabilities of the network were derived by the Bayesian belief network using a small sample of traces. These traces were subjectively interpreted. In principle, the start of the process can be based on a complete lack of any knowledge. In that case, uniform distributions would be used for the conditional probabilities, implying that every outcome is equally probable. Hence over time the system would learn to some degree of accuracy; however such a strategy is not optimal in reducing respondent burden, it is better to start with a learned network.

The imputed data are fed into a Web-based prompted recall survey. Respondents can check the correctness of the imputed activity-travel patterns and revised any incorrect data. This new evidence is then fed back into the Bayesian belief network to update the conditional probabilities. The imputation of activity types follows similar principles but also takes land use information into account. In case such information is not available, the information provided by the respondents on land use is used to dynamically build up a geo-references data base of land uses.

The confirmed activity agendas were viewed as new evidence for the system. These validated activity-travel patterns were used to update the conditional probabilities in the Bayesian belief network. As discussed, it was expected that the accuracy of the system in correctly identifying facets of activity-travel patterns will increase over time with more respondents and traces processed.

# EMPIRICAL APPLICATION

The results of previous pilot study conducted by Moiseeva, Jessurun and Timmermans (Moiseeva et al, 2010) showed promising results: (1) the imputation accuracy for various facets of activity-travel patterns is considerable high and varies between 85-99%; (2) the network learns over time, leading to improved accuracy. The aim of this pilot study is to test alternative learning processes that might lead to obtain the highest imputation accuracy over shorter period of time.

During the multi-day data collection the history of every respondent is collected. This implies that learning can be based on the continuously updated conditional probability tables, aggregated across respondents, or on the personal histories of respondents or on a combination of both. As discussed, it was expected that the imputation accuracy of the system in identifying different facets of activity-travel patterns from GPS tracers will increase over time with more respondents and traces processed.

Further the paper discussed the results of these two alternative approaches to impute transport modes and activities from multi-week activity travel diaries. The personal histories approach is based on *creating personalised Bayesian belief network* for every participant and than learning the conditional probability table of Bayesian network with individual's data obtained from previously processed traces. The aggregated histories approach is based on updated conditional probability table of *common BBN* using previous multi-dimensional sequences of activity-travel patterns aggregated across respondents.

For both approaches activity-travel patterns corrected and confirmed by respondents using Web-based prompted recall instrument have been used to update the conditional probabilities of initial Bayesian network. This means that over time the interpretation of the GPS traces of every respondent is based on the aggregate conditional probability tables that the system learned on the basis of the previously processes cases of this respondent (first approach) or cases aggregated across respondents (second approach).

## Data

To test these alternative approaches for network learning data were collected during 10 weeks study. 5 participants were involved in data collection carrying the GPS logger Bluetooth A+. Every week they upload GPS traces to the web application and confirmed activity-travel diaries.

Even being aware that activities and trips confirmed by respondents might be inaccurate due to personal factor, in this study it was decided to classify corrected and confirmed by respondents the frequencies of trips, varies transport modes and activities over 15 days as a correct or an accurate observation. Thus in order to asses the possible improvements over time, the imputation accuracy of activities and transport modes confirmed by respondents has been compared with new results from the same activity-travel patterns which imputation was based on the updated conditional probabilities tables of BBN from previous cases.

The conditional probabilities were updated once every 14 days during 10 weeks (Figure 1). It means that GPS tracers of the first and second weeks of data collection (input block) were used to updated probabilities and then third and forth weeks GPS tracers were processed again. Further GPS tracers of the third and forth weeks were used as new evidence for updating conditional probabilities tables and traces of fifth and six weeks were interpreted again on the base on new probabilities, and so force. Hence over time new evidences accumulated in the Bayesian belief network. The new results from GPS traces were compared with results confirmed by participants via a prompted recall instrument.

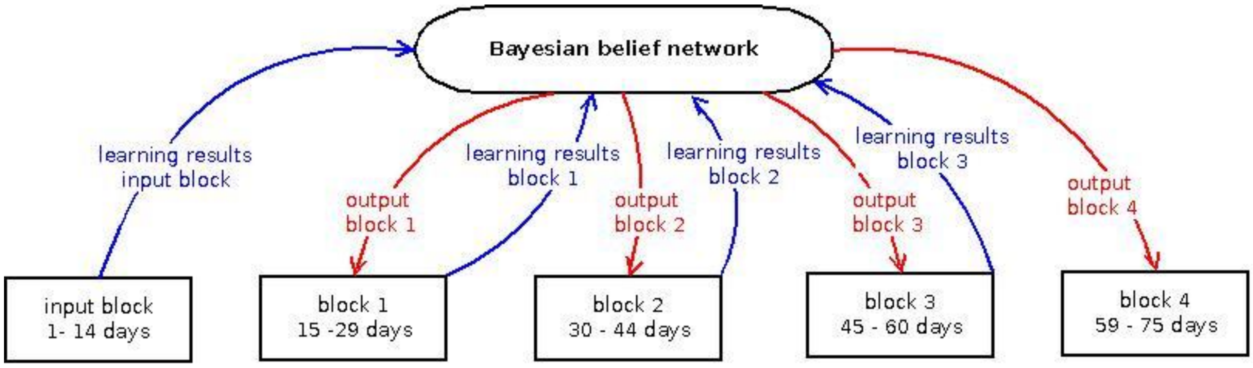


Figure 1 – Learning Scheme

For representation of the learning process over time, weeks have been organised into so-called 4 learning blocks, each block includes 14 days. The first two weeks for every participant form a so-called input block. The input block is the initial block, thus the results of the input block remain the same. In order to show the learning process over time the results form the input block have not been included for calculating the average imputation accuracy of activities and trips.

## Results and discussion

### *Activities*

Table I shows the number of correctly identified activities by the system before learning. During the 10 weeks time period every 15 days in average 151 activities were conducted, 96% of activities were correctly identified by the system. After learning network on the base of personal histories (Table II) and aggregated histories (Table III) correspondently in average 96% and 99% activities were identified. This finding suggest that although the initial network was already quite successful in correctly identifying activity types, the inclusion of the learning algorithm further improved the results.

Table IV and Figure 2 represent the results of the learning process over time. It is remarkable that in case of the aggregated histories learning the imputation accuracy already after weeks increased up to 98% while the imputation accuracy in case of the personal histories learning reached only 94%. This improvement goes steadily over the time. Thus, in this particular case, apparently the similarity between the respondents (traces) is such that the use of aggregate data improves the accuracy of the imputation faster. Note that this should not be necessarily the case. It may also be that inter-subject differences are too big to use the aggregate data successfully.

Table I – Correctly Derived Activities (before learning – confirmed by respondents)

| **Learning Blocks** | **Time** | **Total Records** | **Correctly Derived Activities** | | **Edited Records** | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | **from walking to activity** | **from biking to activity** | **total** | |
| N | days | N | N | % | N | N | N | % |
| input block | 1-14 | 206 | 182 | 88 | 12 | 2 | 14 | 12 |
| 1 block | 15-29 | 171 | 158 | 92 | 11 | 2 | 13 | 8 |
| 2 block | 30-44 | 158 | 153 | 97 | 5 | 0 | 5 | 3 |
| 3 block | 45-59 | 124 | 123 | 99 | 1 | 0 | 1 | 1 |
| 4 block | 60-75 | 150 | 145 | 97 | 4 | 1 | 5 | 3 |
| **Blocks 1-4** | 60 | 603 | 579 | | 21 | 3 | 24 | |
| **Average blocks 1-4** | | **151** | **145** | **96** | | | | |

Table II – Correctly Derived Activities (after learning based on the personal histories)

| **Learning Blocks** | **Time** | **Total Records** | **Correctly Derived Activities** | | **Edited Records** | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | **from walking to activity** | **from biking to activity** | **total** | |
| N | days | N | N | % | N | N | N | % |
| input block | 1-14 | 206 | 182 | 88 | 12 | 2 | 14 | 12 |
| 1 block | 15-29 | 171 | 161 | 94 | 10 | 0 | 10 | 6 |
| 2 block | 30-44 | 158 | 154 | 97 | 4 | 0 | 4 | 3 |
| 3 block | 45-59 | 124 | 121 | 98 | 2 | 1 | 3 | 2 |
| 4 block | 60-75 | 150 | 145 | 97 | 5 | 0 | 5 | 3 |
| **Blocks 1-4** | 60 | 603 | 581 | | 21 | 1 | 22 | |
| **Average blocks 1-4** | | **151** | **145** | **96** | | | | |

Table III – Correctly Derived Activities (after learning based on the aggregated histories across participants)

| **Learning Blocks** | **Time** | **Total Records** | **Correctly Derived Activities** | | **Edited Records** | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | **from walking to activity** | **from biking to activity** | **total** | |
| N | days | N | N | % | N | N | N | % |
| input block | 1-14 | 206 | 182 | 88 | 12 | 2 | 14 | 12 |
| 1 block | 15-29 | 171 | 167 | 98 | 3 | 0 | 3 | 2 |
| 2 block | 30-44 | 158 | 155 | 98 | 3 | 0 | 3 | 2 |
| 3 block | 45-59 | 124 | 123 | 99 | 1 | 0 | 1 | 1 |
| 4 block | 60-75 | 150 | 149 | 99 | 1 | 0 | 1 | 1 |
| **Blocks 1-4** | 60 | 603 | 594 | | 8 | 0 | 8 | |
| **Average blocks 1-4** | | **151** | **149** | **99** | | | | |

Table IV – Correctly Derived Activities (before and after learning)

| Block | time | Activities | | |
|---|---|---|---|---|
| | | Before Learning | Learning Personal Histories | Learning Aggregated Histories |
| N | days | % | | |
| input block | 1-14 | 88 | 88 | 88 |
| 1 block | 15-29 | 92 | 94 | 98 |
| 2 block | 30-44 | 97 | 97 | 98 |
| 3 block | 45-59 | 99 | 98 | 99 |
| 4 block | 60-75 | 97 | 97 | 99 |

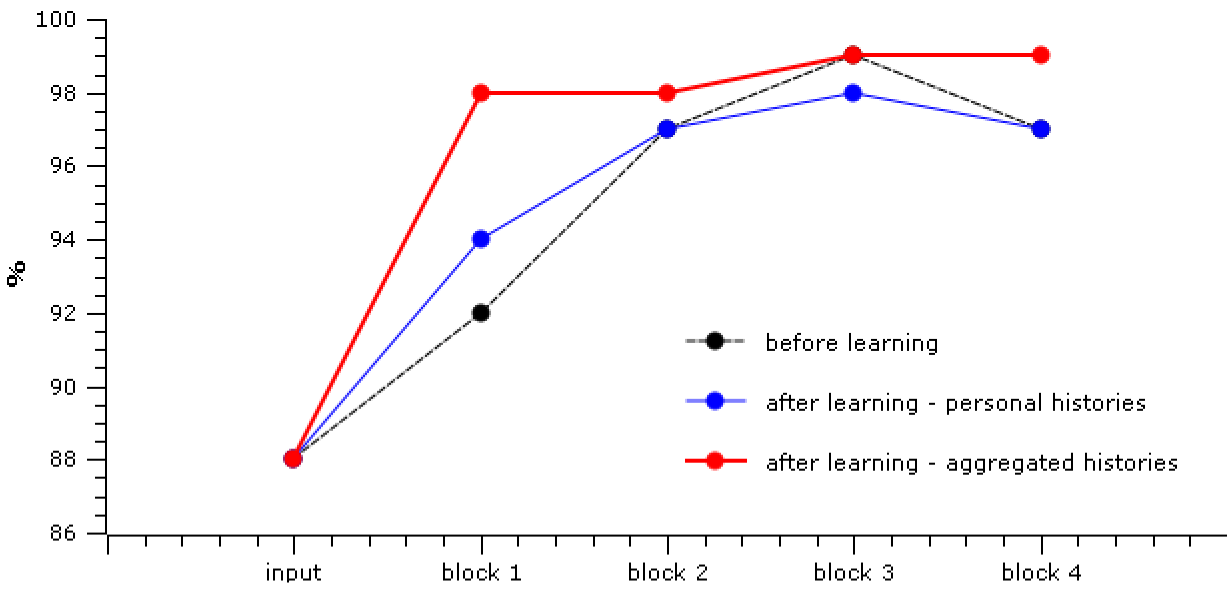


Figure 2 – Imputation of Activities - Learning over Time

### *Trips*

The same learning tendency characterises the imputation accuracy of the all trips. Tables V, VI and VII show that the average accuracy of all correctly identified trips increased from 97% for the personal histories learning and up to 99% for the aggregated histories learning.

Table V – Correctly Derived Trips (before learning – confirmed by respondents)

| **Block** | **Total Records identified as a trip** | **Correct Records from trip to trip** | | **Edited Records from activity to trip** | |
|---|---|---|---|---|---|
| N | N | N | % | N | % |
| input block | 260 | 252 | 97 | 8 | 3 |
| 1 block | 226 | 219 | 97 | 7 | 3 |
| 2 block | 242 | 233 | 96 | 9 | 4 |
| 3 block | 196 | 191 | 97 | 5 | 3 |
| 4 block | 209 | 202 | 97 | 7 | 3 |
| Total block 1- 4 | 873 | 845 | | 28 | |
| **Average blocks 1-4** | **218** | **211** | **97** | | |

Table VI - Correctly Derived Trips (after learning based on the personal histories

| **Block** | **Total Recods identified as a trip** | **Correct Records from trip to trip** | | **Edited Records from activity to trip** | |
|---|---|---|---|---|---|
| N | N | N | % | N | % |
| input block | 260 | 252 | 97 | 8 | 3 |
| 1 block | 226 | 219 | 97 | 7 | 3 |
| 2 block | 240 | 233 | 97 | 7 | 3 |
| 3 block | 196 | 191 | 97 | 5 | 3 |
| 4 block | 211 | 202 | 96 | 9 | 4 |
| Total block 1- 4 | 873 | 845 | | 28 | |
| **Average blocks 1-4** | **218** | **211** | **97** | | |

Table VII - Correctly Derived Trips (after learning based on the aggregated histories across participants)

| **Block** | **Total Records identified as a trip** | **Correct Records from trip to trip** | | **Edited Records from activity to trip** | |
|---|---|---|---|---|---|
| N | N | N | % | N | % |
| input block | 260 | 252 | 97 | 8 | 3 |
| 1 block | 225 | 219 | 97 | 6 | 3 |
| 2 block | 237 | 233 | 98 | 4 | 2 |
| 3 block | 192 | 191 | 99 | 1 | 1 |
| 4 block | 203 | 202 | 100 | 1 | 0 |
| Total block 1- 4 | 857 | 845 | | 12 | |
| **Average blocks 1-4** | **214** | **211** | **99** | | |

### *Transport Modes*

Tables IX - XIII display the results for specific transport modes. These analyses were conducted to examine to what the use of aggregate versus personal data may differ for different transport modes. The results suggest that the system indeed learns fast for all transport modes in the sense that the interpretation accuracy improves over time, especially when learning is based on the aggregated histories across respondents (Table VIII and Figure 3). The imputation accuracy in case of personal histories learning reached max 96% whereas for the aggregated histories learning for last weeks all transport modes were identified with 99 -100% accuracy.

Table VIII – Correctly Derived Modes (before and after learning)

| **Block** | **Time** | **All Modes** | | |
|---|---|---|---|---|
| | | **Before Learning** | **Learning Personal Histories** | **Learning Aggregated Histories** |
| N | days | % | | |
| input block | 1-14 | 88 | 88 | 88 |
| 1 block | 15-29 | 86 | 92 | 96 |
| 2 block | 30-44 | 91 | 96 | 97 |
| 3 block | 45-59 | 96 | 96 | 99 |
| 4 block | 60-75 | 94 | 96 | 100 |

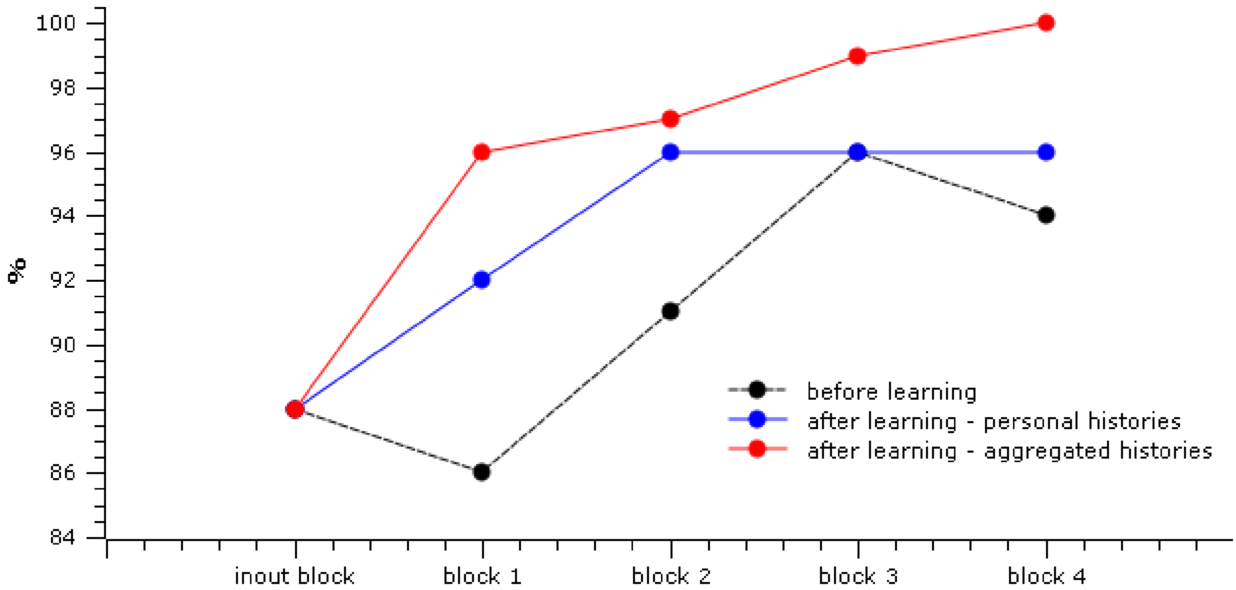


Figure 3 – Imputation of Transport Modes - Learning over Time

Surprisingly the imputation accuracy for results before learning also shows the positive tendency over time. This fact might be explained by the trips distribution over a period of time. For instance, the bus trips, that initially have a lower imputation accuracy in comparison with other types of transport modes (Table IX, XI), have a high frequencies in the first (24 trips) and second blocks (23 trips) while in the third and forth blocks there are significantly less bus trip present (8 trips).

The aggregated histories learning showed better results especially in relation to the public transport modes such as bus and bike (Tables IX, X; XI, XII, XIII). However, it turned out that these differences as depend on the original network. For example, Table XI shows that the accuracy of correctly identifying bus trips increased from 38 % up to 100% for aggregative histories and up to 85% for the personal histories. This high increase in accuracy may be explained by the fact that in the original data very few bus data were present and that initially the system identified the majority of bus trips as train trips. Similarly, a high number of biking trips was first identified as running trips. These errors disappeared very quickly when the system learned new evidences from aggregated data across respondent histories.

Table IX – Correctly Derived Bus Trips (before and after learning)

| **Block** | **Time** | **Bus** | | |
|---|---|---|---|---|
| | | **Before Learning** | **Learning Personal Histories** | **Learning Aggregated Histories** |
| N | days | % | | |
| input block | 1-14 | 63 | 63 | 63 |
| 1 block | 15-29 | 26 | 30 | 100 |
| 2 block | 30-44 | 50 | 100 | 100 |
| 3 block | 45-59 | 50 | 100 | 100 |
| 4 block | 60-75 | 46 | 100 | 100 |
| **Total 1-4** | **60** | **38** | **85** | **100** |

Table X – Correctly Derived Bike Trips (before and after learning)

| **Block** | **Time** | **Bike** | | |
|---|---|---|---|---|
| | | **Before Learning** | **Learning Personal Histories** | **Learning Aggregated Histories** |
| N | days | % | | |
| input block | 1-14 | 85 | 85 | 85 |
| 1 block | 15-29 | 94 | 98 | 97 |
| 2 block | 30-44 | 93 | 98 | 99 |
| 3 block | 45-59 | 92 | 98 | 99 |
| 4 block | 60-75 | 98 | 100 | 100 |
| **Total 1-4** | **60** | **94** | **98** | **99** |

Table XI – Correctly Derived Transport Modes (before learning – confirmed by respondents)

| Learning Blocks | Time | Mode | Total Records | Correctly Derived Modes | | Edited Records | |
|---|---|---|---|---|---|---|---|
| N | days | | N | N | % | N | % |
| input block | 1-14 | walking | 141 | 131 | 93 | 10 | 7 |
| | | bike | 78 | 66 | 85 | 13 | 15 |
| | | bus | 24 | 15 | 63 | 9 | 36 |
| | | train | 9 | 9 | 100 | 0 | 0 |
| | | **TOTAL** | **252** | **221** | **88** | **32** | **12** |
| | | | | | | | |
| 1 block | 15-29 | walking | 96 | 89 | 93 | 7 | 7 |
| | | bike | 95 | 89 | 94 | 13 | 6 |
| | | bus | 23 | 6 | 26 | 17 | 74 |
| | | train | 7 | 7 | 100 | 0 | 0 |
| | | **TOTAL** | **221** | **191** | **86** | **37** | **14** |
| | | | | | | | |
| 2 block | 30-44 | walking | 96 | 87 | 91 | 9 | 9 |
| | | bike | 123 | 115 | 93 | 8 | 7 |
| | | bus | 8 | 4 | 50 | 4 | 50 |
| | | train | 6 | 6 | 100 | 0 | 0 |
| | | **TOTAL** | **233** | **212** | **91** | **21** | **9** |
| | | | | | | | |
| 3 block | 45-59 | walking | 93 | 91 | 98 | 2 | 2 |
| | | bike | 93 | 91 | 98 | 2 | 2 |
| | | bus | 8 | 4 | 50 | 4 | 50 |
| | | train | 3 | 3 | 100 | 0 | 0 |
| | | **TOTAL** | **197** | **189** | **96** | **8** | **4** |
| | | | | | | | |
| 4 block | 60-75 | walking | 82 | 79 | 96 | 3 | 4 |
| | | bike | 104 | 102 | 98 | 2 | 2 |
| | | bus | 13 | 6 | 46 | 7 | 54 |
| | | train | 3 | 3 | 100 | 0 | 0 |
| | | **TOTAL** | **202** | **190** | **94** | **12** | **6** |

Table XII – Correctly Derived Transport Modes (after learning based on the personal histories)

| Learning Blocks | Time | Mode | Total Records | Correctly Derived Mode | | Edited Records | |
|---|---|---|---|---|---|---|---|
| N | days | | N | N | % | N | % |
| input block | 1-14 | walking | 141 | 131 | 93 | 10 | 7 |
| | | bike | 78 | 66 | 85 | 13 | 15 |
| | | bus | 24 | 15 | 63 | 9 | 36 |
| | | train | 9 | 9 | 100 | 0 | 0 |
| | | **TOTAL** | **252** | **221** | **88** | **32** | **12** |
| | | | | | | | |
| 1 block | 15-29 | walking | 96 | 87 | 91 | 9 | 9 |
| | | bike | 95 | 93 | 98 | 2 | 2 |
| | | bus | 23 | 16 | 30 | 7 | 70 |
| | | train | 7 | 7 | 100 | 0 | 0 |
| | | **TOTAL** | **221** | **203** | **92** | **18** | **8** |
| | | | | | | | |
| 2 block | 30-44 | walking | 96 | 89 | 93 | 7 | 7 |
| | | bike | 123 | 120 | 98 | 3 | 2 |
| | | bus | 8 | 8 | 100 | 0 | 0 |
| | | train | 6 | 6 | 100 | 0 | 0 |
| | | **TOTAL** | **233** | **223** | **96** | **10** | **4** |
| | | | | | | | |
| 3 block | 45-59 | walking | 128 | 123 | 96 | 5 | 4 |
| | | bike | 88 | 86 | 98 | 3 | 2 |
| | | bus | 8 | 7 | 100 | 1 | 0 |
| | | train | 3 | 3 | 100 | 0 | 0 |
| | | **TOTAL** | **227** | **219** | **96** | **9** | **4** |
| | | | | | | | |
| 4 block | 60-75 | walking | 82 | 73 | 89 | 9 | 4 |
| | | bike | 104 | 104 | 100 | 0 | 0 |
| | | bus | 13 | 13 | 100 | 0 | 0 |
| | | train | 3 | 3 | 100 | 0 | 0 |
| | | **TOTAL** | **202** | **193** | **96** | **9** | **4** |

Table XIII – Correctly Derived Transport Modes (after learning based on the aggregated histories across participants)

| Learning Blocks | Time | Mode | Total Records | Correctly Derived Mode | | Edited Records | |
|---|---|---|---|---|---|---|---|
| N | days | | N | N | % | N | % |
| input block | 1-14 | walking | 141 | 131 | 93 | 10 | 7 |
| | | bike | 78 | 66 | 85 | 13 | 15 |
| | | bus | 24 | 15 | 63 | 9 | 36 |
| | | train | 9 | 9 | 100 | 0 | 0 |
| | | **TOTAL** | **252** | **221** | **88** | **32** | **12** |
| | | | | | | | |
| 1 block | 15-29 | walking | 94 | 89 | 95 | 5 | 5 |
| | | bike | 95 | 92 | 97 | 3 | 3 |
| | | bus | 23 | 23 | 100 | 0 | 0 |
| | | train | 7 | 7 | 100 | 0 | 0 |
| | | **TOTAL** | **219** | **211** | **96** | **8** | **4** |
| | | | | | | | |
| 2 block | 30-44 | walking | 96 | 90 | 94 | 6 | 6 |
| | | bike | 123 | 123 | 99 | 1 | 1 |
| | | bus | 8 | 8 | 100 | 0 | 0 |
| | | train | 6 | 6 | 100 | 0 | 0 |
| | | **TOTAL** | **233** | **227** | **97** | **7** | **3** |
| | | | | | | | |
| 3 block | 45-59 | walking | 94 | 93 | 99 | 1 | 1 |
| | | bike | 86 | 85 | 99 | 1 | 1 |
| | | bus | 8 | 8 | 100 | 0 | 0 |
| | | train | 3 | 3 | 100 | 0 | 0 |
| | | **TOTAL** | **191** | **189** | **99** | **2** | **1** |
| | | | | | | | |
| 4 block | 60-75 | walking | 82 | 81 | 99 | 1 | 1 |
| | | bike | 104 | 104 | 100 | 0 | 0 |
| | | bus | 13 | 13 | 100 | 0 | 0 |
| | | train | 3 | 3 | 100 | 0 | 0 |
| | | **TOTAL** | **202** | **201** | **100** | **1** | **0** |

# CONCLUSIONS

Although the relevance and potential of GPS traces for collecting activity-travel data, especially over a longer period of time has been argued in the transportation research community and beyond, and some authors pointed at the possibility of using learning algorithms (Auld et al, 2009, Rudloff and Ray, 2010), an examination of the relevant literature in transportation research suggest that the actual application of learning algorithm has not yet been widely realised. The success of any learning algorithm depends on how it is implemented. In the present study, we compared the use of personal versus aggregate histories in updating a Bayesian belief network that serves as the learning algorithm. Confirmed and corrected imputations of elements of an activity-travel pattern are viewed as new evidence in this approach, and the conditional probability tables are systematically uploaded. This implies that the adequacy of aggregated versus personal histories depends on the degree to which the activity-travel profile of an individual respondents resembles the aggregate distribution. If this is the case, the use of aggregate histories will imply a faster learning; if not, the aggregate distribution may be atypical and therefore misclassify the behavior of an individual traveler.

The results of the present still small-scale study suggest that the sample is sufficiently homogeneous to use the aggregate histories.  The results indicate the learning is higher when aggregate histories are used, at least for this sample. Moreover, results suggests that if for whatever reason, the initial rules or principle used by the researchers are wrong or not very effective, the system will identify and correct these changes fast.

Some final comments seem in order. First, the results also indicate that sometimes the use of personal histories produces better results. This will happen in case the profile of an individual traveler differs from the aggregate distribution.  Second. There is no guarantee that the application of the learning algorithm will necessarily improve over time as some results suggest. One of the reasons may be a systematic shift in the relationship between the classification and the condition variables. It implies that in case of longer period data collection, it may be advised to apply structure learning such that the Bayesian network is optimized in light of the accumulated data. Thirdly, it should be emphasized that the results reported in this paper, depend on the kind of approaches compared. Both the aggregated and personal histories use the Bayesian belief network. This implies that the system learns probabilistically. An alternative may for instance be to use a deterministic personal approach: what was the behaviour, the last time under the same set of conditions.

In any case, and although such differences in approaches will generate different findings, the results of the present study do provide evidence for the claim that learning algorithms may be powerful tools in reducing respondent burden in multi-week surveys of activity-travel behaviour. It seems time for large-scale applications.

# REFERENCES

Auld, J., Williams, C., Mohammadian, A., and P. Nelson. (2009) An automated GPS-based prompted recall survey with learning algorithms. *Transportation Letters*, 1, 58-79.

Axhausen, K.W., S. Schönfelder, J. Wolf, M. Oliveira and U. Samaga. (2004) Eighty Weeks of Global Positioning System Traces: Approaches to Enriching Trip Information, In Transportation Research Record: Journal of the Transportation Research Board, No. 1870, Transportation Research Board of the National Academies, Washington D.C., 46-54

Bohte, W., and K. Maat. (2008) Deriving and validating trip destinations and modes for multi-day GPS-based travel surveys: An application in the Netherlands. Presented 87th Annual Meeting of the Transportation Research Board, Washington D.C.

Chung, E. and A. Shalaby. (2005) Development of a Trip Reconstruction Tool for GPS-Based Personal Travel Surveys, Journal of Transportation Planning and Technology, 28, 381-401.

Doherty, S.T., Lee-Gosselin, M.E.H., and D. Papinski. (2006) Internet-based prompted recall diary with automated GPS activity-trip detection: system design. Presented at 85Th Annual Conference of the Transportation Research Board, Washington, D.C.

Li, Z. and A. Shalaby. (2008) Web-Based GIS System for Prompted Recall of GPS-Assisted Personal Travel Surveys: System Development and Experimental Study. Presented at 87th Annual Meeting of the Transportation Research Board, Washington D.C., USA.

Marca, J., C.R. Rindt and M. McNally. (2002) Collecting activity data from GPS readings. Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., USA.

Moiseeva, A., Jessuran, J and Timmermans, H.J.P. (2010) Semi-automatic imputation of activity-travel diaries using GPS traces, prompted recall and context-sensitive learning algorithms, To appear in Transportation Research Record: Journal of the Transportation Research Board

Rudloff, C. and M. Ray. (2010) Detecting travel modes and profiling commuter habits solely based on GPS data. Presented at 89st Annual Meeting of the Transportation Research Board, Washington D.C., USA.

Stopher, P., and A. Collins. (2005) Conducting a GPS prompted recall survey over the internet. Presented 84th Annual Meeting of the Transportation Research Board, Washington, D.C.

Timmermans, H.J.P., Moiseeva, A., Jessurun, J., (2009) Collecting Activity-Travel Data Using Modern ICT: Recent developments and Prospects, Proceedings 14th HKSTS International Conference