Preface

This master thesis is the last step to achieve a master degree at the University of Hasselt, Belgium in the field of transportation sciences, specialization mobility management. I would like to thank my promotor, prof. dr. Davy Janssens to give me the opportunity to graduate under his supervision with one of his topics. I also like to thank very much my supervisor, ir. Wim Ectors, who guided me through the whole year, taught me how to model data and gave me insight in what scientific research really is. Of course, I thank my family and my parents for the opportunity they gave me to study and anyone who stirred my interests to finish this thesis.

Summary

This thesis researches the National Household Travel Survey which was made in the United States of America. This rich database contains information about personal, household and mobility characteristics. The aim of the research is to identify which personal and household variables influence mobility outcomes and to quantify this. Literature was explored to identify which characteristics. Afterwards these characteristics were checked using decision trees (the C4.5 algorithm). Finally, these correlations were quantified in regression and discrete choice models. This research constructed four models, predicting tour type, activity duration, traveled distance and mode choice.

In the database, tour types can begin or end at home, at work or any other place (called others). This is because many tours are work related. The tour type is influenced the strongest by the begin and end time of the tour, whether it is in a weekend or not, if it is carried out by active, non-active people or by people that go to school, the distance of the longest segment and the great circle distance to work.

The mode choice model predicts the probabilities of using public transport, bicycling or walking as mode choice with car use as a reference. It is influenced the strongest by the composition of the household (bachelor, children, retired …), educational level, whether there were intermediate stops or not, the total traveled distance, the number of vehicles in the household, the housing density of the neighborhood and the begin time of the trip.

Activity duration was the most affected by the travel time, the age of an agent, whether the agent can adjust his or her begin and end times at work, whether the activity is in a weekday or not, the time of the day, the type of worker (part time, fulltime or having multiple jobs) and by gender.

Traveled distance is mostly affected by distance to work, the respondent living in an urban or rural environment, the time of day, whether it is during the weekend or not and some environmental characteristics (population density, renter occupancy and working people per square mile).

After these models had been constructed, they were compared and integrated with each other. This part is the conclusion and holds a discussion for further research.

Contents

1. Introduction: activity based modelling

Traffic can be seen as a deduced demand of the desire to perform activities. In this view, the reasons for traveling are captured rather than the final trips to predict travel demand. Models have been built to describe current and hence, future traffic situations based on activities of respondents (Chandra R. Bhat, 1999).

This data is captured by several surveys in several countries ((National Household Travel Survey) ,(D. Janssens, 2013),(Centraal Bureau Statistiek)).

- "OVG" ("Onderzoek Verplaatsingsgedrag", Research transportation behavior) in Flanders.
- "OVIN" ("Onderzoek Verplaatsingen In Nederland", Research transportation in The Netherlands) in The Netherlands
- NHTS ("National Household Travel Survey") in the United States
- …

This data about activities is an input for activity based modelling. Also, there is an interaction between the trips and the (adaptation of) activities. Figure 1 visualizes this (McNally, 2000).

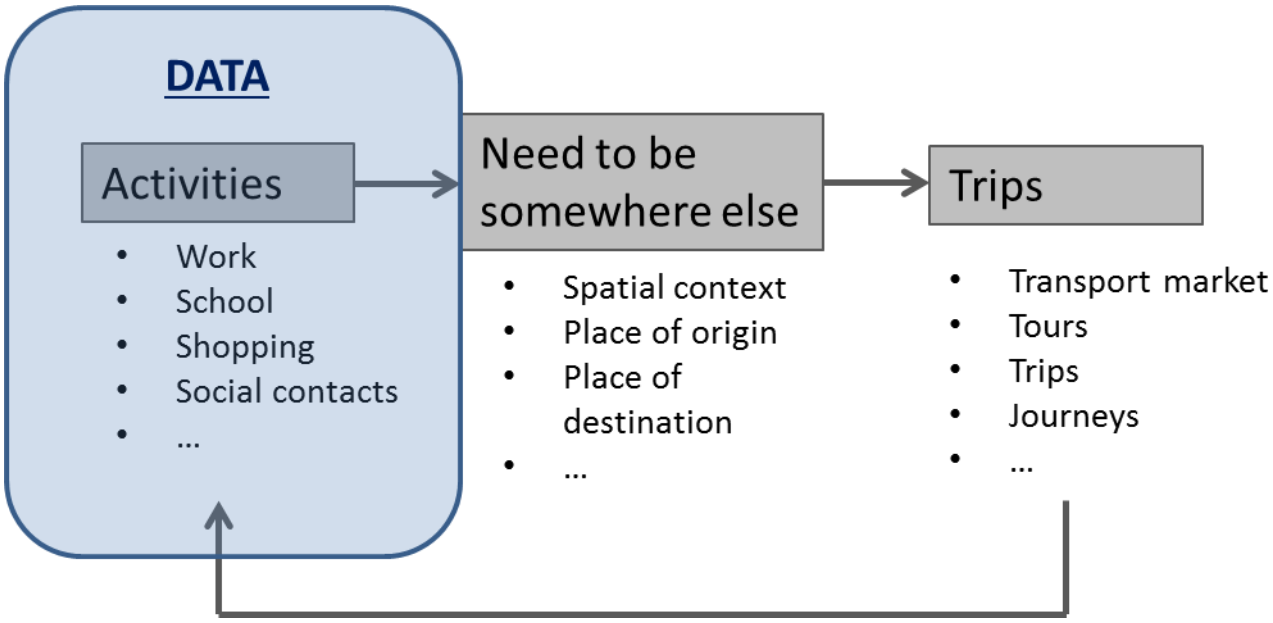

Figure 1: activity based models

2. Aim of the research

The data described in the section above is the input of a model and hence, it defines the accuracy of all its predictions. This data can achieve a great magnitude. Problems that might arise when capturing data are shown below (William Davidson, 2006).

- Reported activity duration: the time it takes to perform an activity can be wrong (e.g. overlapping) due to the inability of respondents to recapture these details, reporting mistakes, …
- Household joint activity and travel: the interactions between members of a household seem to be often wrongly reported (mistakes, one member is more accurate in responding than another, …). Technology can solve this problem partially (e.g. electronic diaries asking respondents how they'll return after going somewhere, …)
- Advanced models may need a broader set of activities than is currently available. E.g. activities performed at home are often excluded (at home working, at home sporting, …), however they can have an effect.
- Underreporting trips by simply forgetting. In home activities and activities leading to short trips are sensitive to this. Again, technology can be a solution.
- The complexness of data makes it difficult to validate the model: are consequences of changed variables due to coincidences, variability typical for the model used or is there really an effect?

Also, capturing data requires a lot of effort. The second OVG Flanders (research travel behavior Flanders) used a random sample of at least 2500 households that were given a questionnaire that had to be filled in with "paper and pencil" or were filled in by researchers contacting the respondents by telephone. From 2008 (OVG 3), they used a stratified sample of 8000 persons contacted face to face or by post. In the OVG 4 the same methods as OVG 3 were used, but the 8000 persons were spread over a period of 5 years (each year at least 1600 persons). The research was done from 2008 until 2013. The methodological differences make comparisons between the results scientifically impossible. Note that since OVG3 there are accurate comparisons possible

The OVIN (Onderzoek Verplaatsingen In Nederland, Research relocations in The Netherlands) took samples among Dutch people and they were asked about their activities for one predefined day: which activity, where was it, how far was the trip, which mode was used, how long took the activity … . This was combined with demographic data of the respondents to describe the travel behavior of Dutch inhabitants. They were contacted with a questionnaire on the internet, by telephone or face-to-face. Each of these methods has its advantages and disadvantages.

In the NHTS (the American National Household Travel Survey), 400000 recruiting forms were sent to households. All types of Americans able to travel were covered. When there were lingual or age restrictions, this was solved by the use of proxy interviews (a person responsible for the original respondent took the interview, see further). 150147 households remained. They were spread over all the 50 states of the USA. Each of these households was assigned automatically a random date for reporting activities. These were the steps for the respondents to ensure they would report reliable data:

1. Letter with a 5,00$ fee and information about the research
2. Additional phone call after one week
3. Diaries for their assigned day, containing (literal citation):
    - "A letter from the U.S. DOT thanking the household for completing the household interview and agreeing to participate in the survey;
    - A brochure describing the survey;

- A travel day diary and a two-dollar cash incentive in individual envelopes personalized for each household member at least five years old. The reverse side of each diary provided guidance on completing the diary and included an example of a completed diary;
- An eye-catching brightly colored reminder card identifying the household's travel date;
- An odometer mileage form identifying the make, model and year of each household vehicle, with spaces to enter the odometer readings and the dates they were taken (National Hosehold Travel Survey, 2011)."

4. Reminder call to ensure if the diaries were delivered and to answer further questions.
5. Within one week, the respondents should have had a phone call of the interviewer to note the data of the diaries. This is where the proxy interviews were used (see above). The operational work was made easier by computer assisted telephone interviewing (CATI).

- To reduce the respondents burden, interviewers were trained and trip rosters were automatically adjusted by them in case of "joint trips" (trips made together by two or more members of the household).

Due to all these efforts, it took more than one year to capture the data (March 2008 until May 2009). Additional efforts were the training of respondents and developing a computer system for CATI (U.S. Department of transportation Federal Highway Administration, 2011).

The aim of this research is to reduce some of the efforts made in capturing data by identifying general characteristics, which can be used in activity based models. One example might be the correlation of age and modal choice or the correlation of gender and trip length distribution. These are examples to illustrate the aim of the research and no statements that there is actually such a correlation!

Figure 2 shows the concept for an activity based model this research tries to achieve.

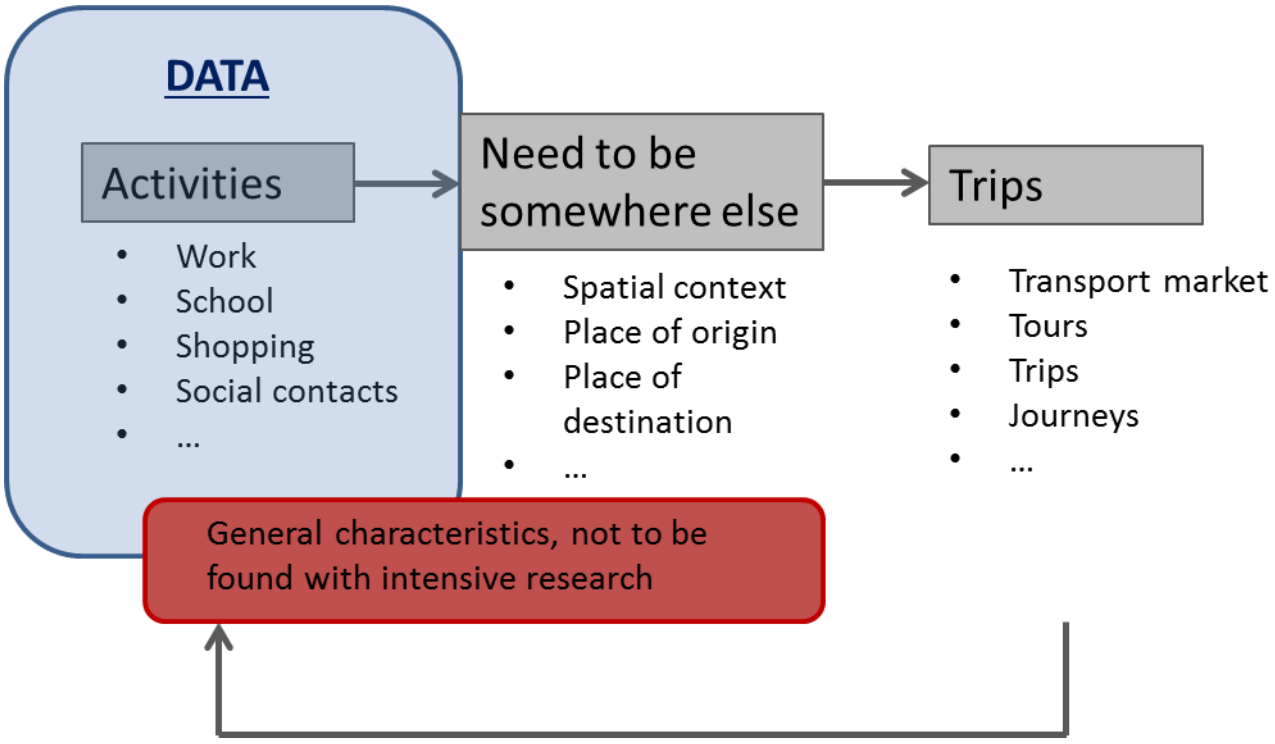

Figure 2: changed activity based modelling

3. Literature Review

This chapter discusses the personal, household and environmental characteristics influencing mobility outcomes (such as mode choice, traveled distance, activity participation …). It is important to note that the information comes from scientific papers and that these may cover different fields (personal, household and environmental characteristics). The same source may come back in different sections.

3.1 Personal characteristics and mobility outcomes

This section describes the effect of variables that are PERSONAL to a person (not to a household). It are variables such as age, income, driving license, gender and so on.

A first research that talks about the influence of sex and mobility outcomes is the Research Travel Behavior Flanders (Onderzoek Verplaatsingsgedrag Vlaanderen, OVG, 2014). They state that sex influence traveled distance (males travel averagely 42.83 kilometers a person a day and females only 32,30 kilometers). They mention that this gap remains stable over the years. This same research delivers information on travel mode choice among both sexes (figure 3).

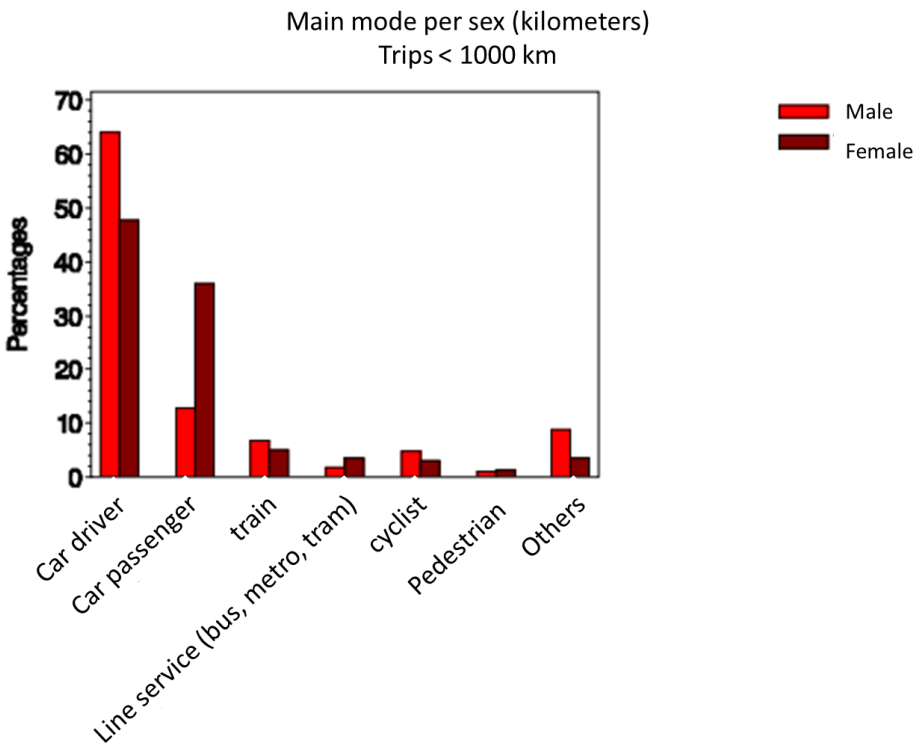

Figure 3: mode distribution among sexes (OVG Vlaanderen, 2014, translated)

It is clear that for both categories, car driver and car passenger are the most frequent modes. There seems to be a huge difference in the percentage a man or a woman is car driver and car passenger. Therefore, in this research there is assumed that sex is an adequate explaining variable of transport mode and especially for the binary dependent mobility variable car passenger or car driver.

Figure 4 was copied (and translated) from the OVG and shows the differences in travel motives between sexes. It becomes clear that sex can be a declaring variable for the appearance of a professional trip (work or business) or for a leisure trip (relaxation, sport, culture …). Males will perform, corresponding to figure 4, more professional trips, and women compensate this with spending more of their kilometers to other motives, although these differences stay smaller.

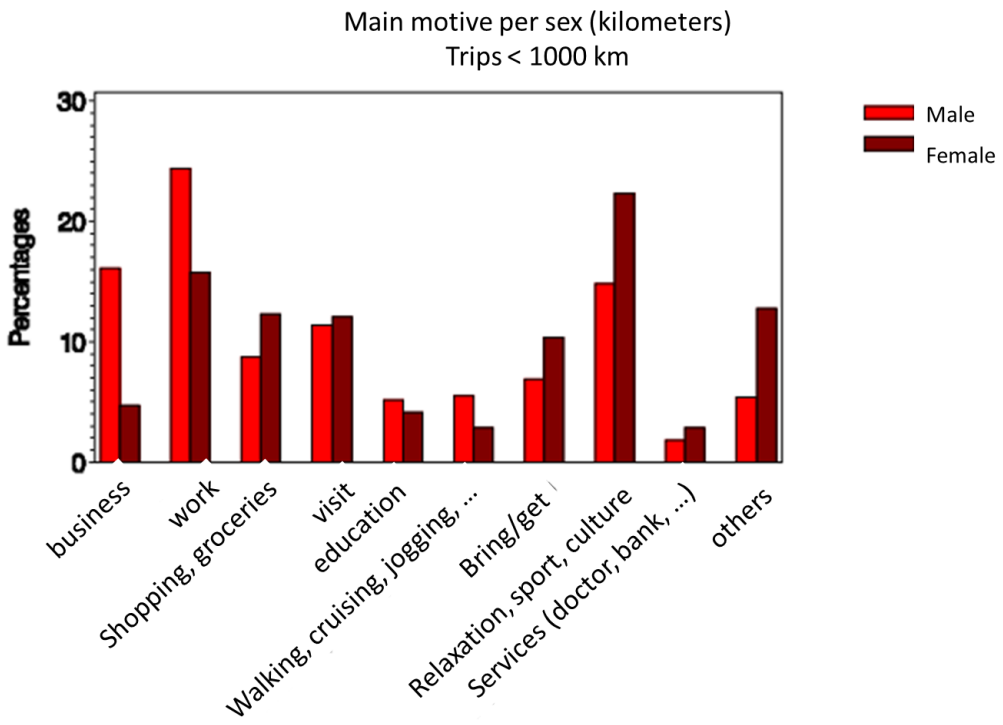

Figure 4: motives and sexes (OVG Vlaanderen 4.5, translated)

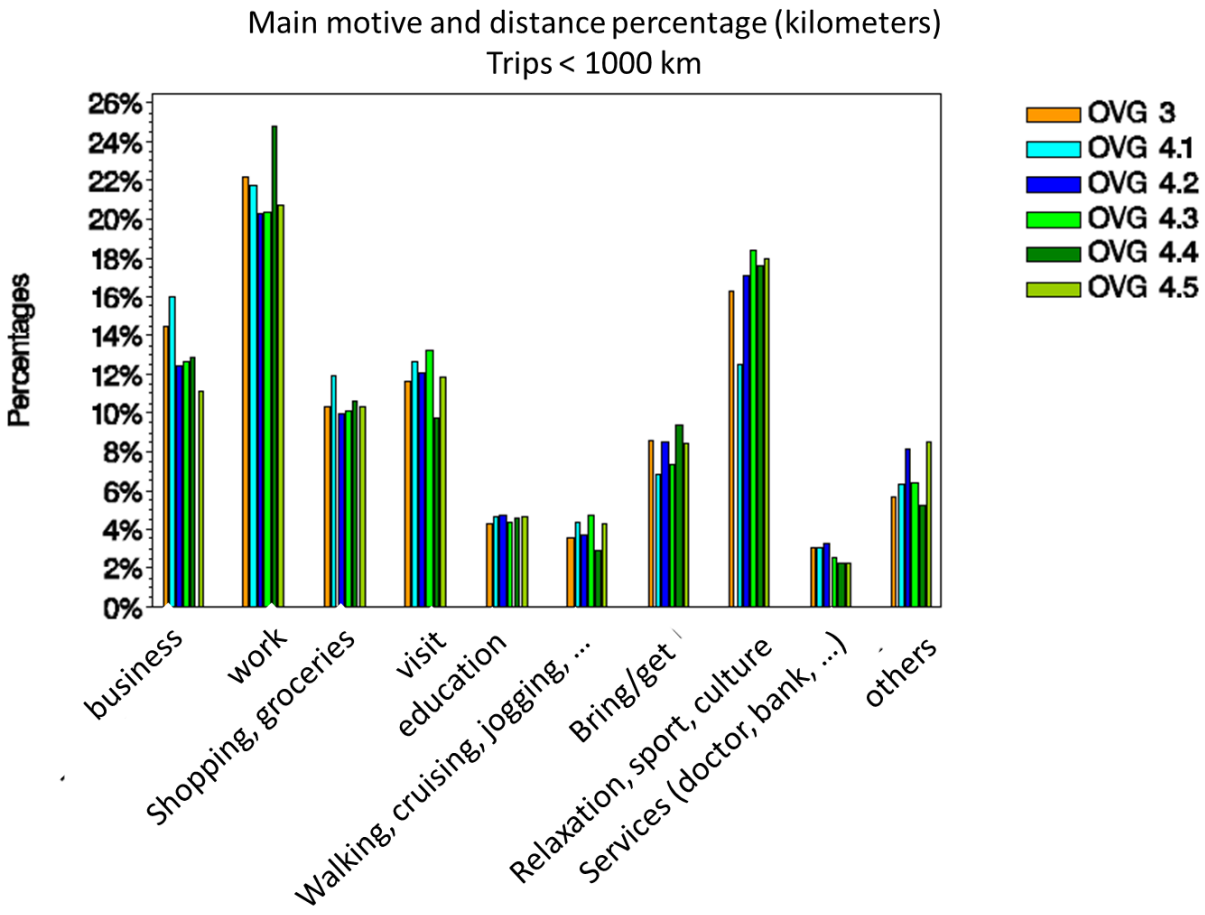

Figure 5: motive and distance percentage

Figure 5 reveals that the three motives that seem to have a significant difference of appearance between both sexes also take the top three of total percentage of traveled kilometers. Therefore, it can be justified to investigate the effect of gender on the appearance of the three motives (separately) and the correlation between their expected traveled kilometers per motive (D. Janssens, 2014).

The influence of sex is also confirmed by the New Danish National Model, especially in combination with age (Jeppe Rich, 2010). Other personal factors are described as well. This source will be recaptured during the next parts of the literature review s it does not only cover personal characteristics. It describes the framework of a Danish activity based model. It has several levels (figure 6). The model assumptions describe the social and physical environment. These are seen as the driving forces of making choices, which are further explained in the strategic model (long term choices, e.g. buying a new house on a specific location). This is the source of transport demand, which gets assigned to a specific mode. The synthesis of the population defines freight transport needs, but the freight transport model isn't discussed, as the research focuses on human characteristics. Also, the model uses an algorithm to compose a population.

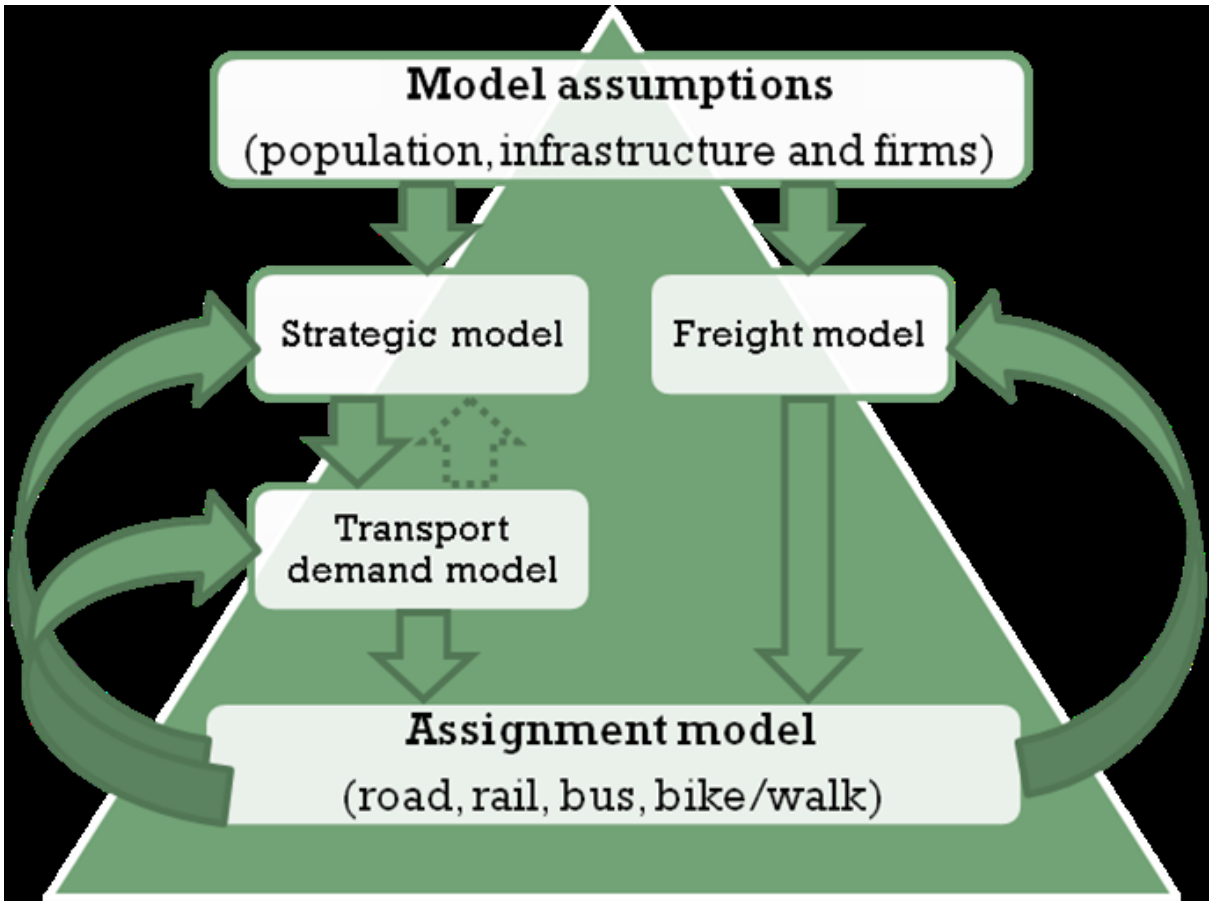

Figure 6: Danish activity model structure

The transport demand model is divided in five sub models, as shown in figure 7. It is fed by the strategic model, as was already explained.

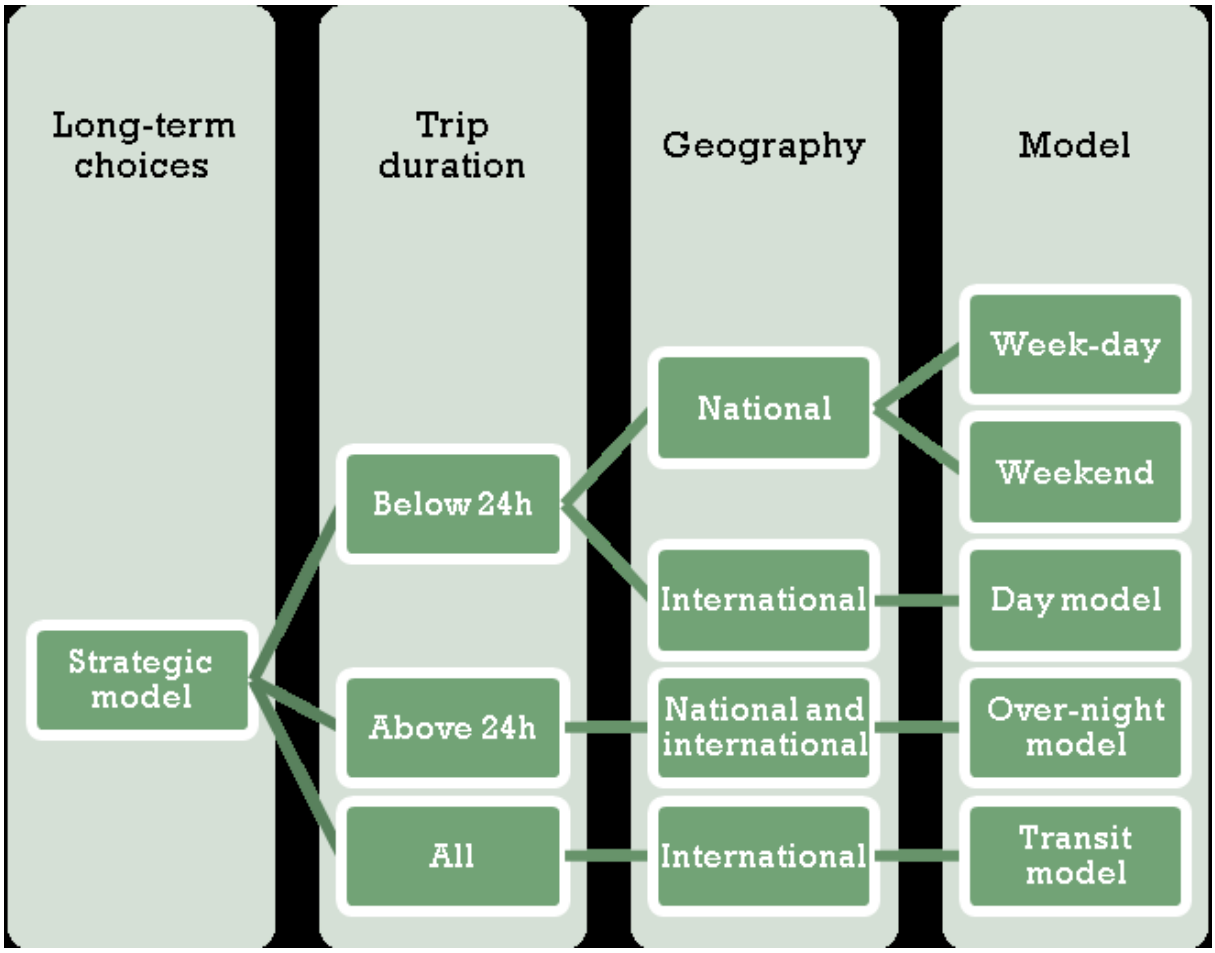

Figure 7: demand model

As can be deduced from the figure above, the model generates decisions on several levels in different categories. These decisions are finally used to make different types of models (see the last step).

First, the synthesizer seems to let the variable “age” have a combined effect with all other variables:

- Age*gender;
- Age*income;
- Age*Lma (labor market area);
- Income*Lma

These variables proved to be good to predict mobility outcomes.

Note: labor market area is an idea proposed by US government and it defines an area in which a resident can find work in an acceptable commuting distance (Maine, 2015).

In 2007, a research was carried out to influence motorcycle usage in Malaysia. This research (Ibrahim Sheikh A. K., 2007) developed a model to predict mode choice behavior of the (unwanted) use of the motorcycle. It was defended because this is a major mode in developing countries (e.g. Laos and

Vietnam) and causes a lot of road safety problems. One of their assumptions is that bus users and motorcyclists have the same income category and hence, it is possible to change their behavior. Travel time to work or school was also said to be an influencing factor. Further, the focus is on mode specific factors rather than personal or household related ones, but they provide a table of their respondents with some demographic factors. In the final model, variables from table 1 were found to be significant.

Constants	Coefficients	S.E.	Sig.	Odd ratio
Age	1.075	0.334	0.001	2.93
Gender	1.921	0.342	0.000	6.8
Travel time	−0.158	0.017	0.000	0.854
Travel cost	−0.047	0.021	0.000	0.954
Income	−0.817	0.353	0.000	0.442
Constant	4.215	0.779	0.000	
(−2)log likelihood		280.158		

Table 1: estimation result for binary mode choice model

For the main research, it is important to note that age, gender and income are personal characteristics influencing mode choice (using female as reference for gender). It is very important to note that this mode choice model is binary! The influence of income may also be different for other countries since it is directly said that this research focuses on developing countries.

To continue with a person's work situation, Linda Nijland (2014) linked time availability to the time one works a week and showed that this has an effect in utility functions for performing different types of activities. There was also shown that the time passed since an activity was last performed also has its effect. The utility function of performing an activity of type "i", is (Linda Nijland, 2014):

$U_{sdi} = V_{sdi} + V_{di} + \varepsilon_{sdi}$

Where:

- "d" is the current day of the week;
- "s" is the day activity of type "i" was conducted the last time before d;
- Vsdi is the need-related utility of activity i built-up between s and d;
- Vdi is a (positive or negative) preference for conducting activity i on day d;
- ε_{sdi} is an error term.

The need related to perform a utility over time is expressed as:

$V_{sdi} = \beta_i \ln(t_i+1)$

Where:

- β_i = a need related utility growth rate;
- $t_i = (d - s)_i$.

These comparisons are the basis of the model that simulates postponement and adaptation behavior when future activities of the same type occur unexpectedly. For this research, the basic model is most interesting as it reveals the influence of time availability on activity performing. Postponement is something that this paper does not consider to be a consequence of socio demographic characteristics and hence, there can't be measured a regression line, which is the main research question of this thesis. However, a description on the basic simulation can provide insights.

The following activities were included in the simulation:

- Shopping–one store;
- Shopping–multiple stores;
- Service-related activities;
- Social activities;
- Leisure activities (other than touring);
- Touring (by car, bike, or foot).

The results of the simulation are in table2:

Simulation results: basic model

	Shopping–one store	Shopping–multiple stores	Service-related activities	Social activities	Leisure activities	Touring (car, bike or foot)
Zero hours work a week						
Freq	35	9	6	16	7	12
n Mon	5	2	2	3	0	4
n Tue	4	0	0	1	0	0
n Wed	5	2	0	4	0	3
n Thu	6	0	2	1	0	1
n Fri	6	2	1	1	0	0
n Sat	8	3	1	4	5	1
n Sun	1	0	0	2	2	3
40 h work a week						
Freq	28	8	4	11	7	10
n Mon	2	0	0	0	0	0
n Tue	3	0	0	0	0	0
n Wed	7	0	0	2	0	0
n Thu	2	0	0	0	0	1
n Fri	5	2	0	0	0	2
n Sat	8	5	4	7	5	3
n Sun	1	1	0	2	2	4
24 h work a week						
Freq	34	12	5	14	7	12
n Mon	3	0	0	0	0	0
n Tue	4	0	0	0	0	0
n Wed	8	4	2	6	0	4
n Thu	1	0	0	0	0	0
n Fri	11	4	2	1	0	3
n Sat	6	3	1	5	5	1
n Sun	1	1	0	2	2	4

Table 2: simulation results (scanned directly from source)

The main conclusions are:

- There is a clear and logic relationship with the parameter time:
  - Time availability
  - Time passed since activity type was performed
  - Day of week
- Part time workers (24h): Wednesday and Friday are preferred for activities
- Full time workers (40h): weekend days are preferred for activities
- There is a particular effect for the activity type of visiting one shop: it is performed mostly on Friday or Saturday.

By investigating the methodology of Bayesian Networks compared to Decision Trees, the variables start time and duration were investigated (Davy Janssens, 2005). This is also a time component important to individuals and their mobility behavior. Renni Anggraine (2010) showed the importance of day of the week.

Shiftan showed in 2008 that employment situation influences the activity behavior of people. This corresponds with the influence of working time. This same research mentions car ownership as an important factor. This is discussed further.

With a research using stated preference and revealed preference data to build a model predicting the public transit share rate (Zhihu Zhanga, 2013), the interviews consisted of three parts: personal information (gender, age, occupation, car purchase plan, and monthly household income), revealed preference of bus use and a stated preference survey. In the stated preference survey, the focus was on variables that can be changed and influence bus use, i.e. variables describing the bus service and no personal variables. For the thesis, these are not important, but the personal variables eventually found significant and included in the model, are. **However, it is very important to remind that this research shows that mode choice is not fully explained by personal variables, but also by mode specific variables such as ticket price, frequency, comfort …**

The research uses an overall multi logit model of the form:

$$P_{in} = \frac{\exp(\theta V_{in})}{\sum_{j=1}^{J} \exp(\theta V_{jn})} \qquad i = 1,2,\dots,J$$

with p being the probability of any alternative i chosen by person n from choice set J , θ is an unknown coefficient and V_{in} is called the systematic components of the utility of alternative i.

To combine the data of the revealed and stated preference survey in the model estimation, mathematical procedures were used that can be checked in the paper. Their final conclusions were about factors specific to bus trips and how users can continue their behavior public transport use or change it. They acknowledge that there were no non-users investigated and hence, a real choice was not modeled. However, this paper suggests influence of **personal characteristics** together with mode specific characteristics on the mode choice as a whole.

3.2 Household characteristics and mobility outcomes

The learning based model ALBATROSS (Theo A. Arentze, 2002) assumes that performing activities is the result of a decision making process. Firstly, long term decisions influence choice behavior (marital status, number of children, residential location, work place, work type and having availability to transport modes). These long term decisions influence socio demographic variables. The learning based part is explained by time pressure of performing activities and time that last activities of a particular type were performed. They also found upon reinforcement and social learning; respectively revealed to be influences from environment and household interactions.

The order of different decision making and is shown in figure 7. This process should result in a daily schedule with combined activities and influence of other decisions already made.

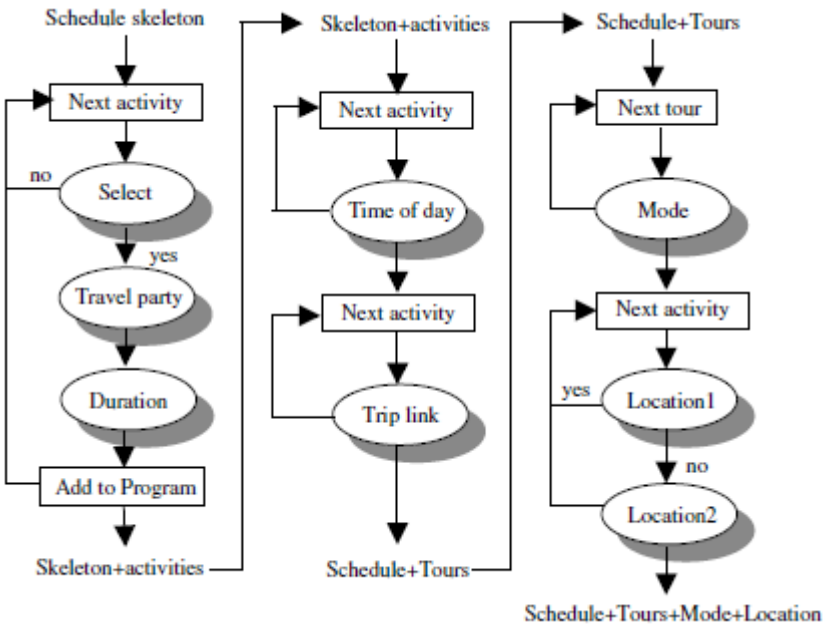

Figure 8: decision making process

The ovals in figure 8 are the points where decision trees give input. These decision trees were developed using a tradeoff between the algorithms C4.5 and CART. The data was collected by diaries in which the days of the week were balanced so that each day had the same frequency. Respondents filled in their activities and corresponding characteristics (time, mode …) from a choice set of 48 activity types. Finally, data of 2198 households remained to build decision trees. This was complemented with environmental data (opening hours, physical constraints …) and as such, the respondents were assigned to a Traffic Analysis Zone.

There were made decision trees for the following characteristics:

- Mode for work;
- Activity type;
- With who is the activity performed;
- Activity duration;
- Activity time of day;
- Trip link;
- Mode for other activity than work;
- Activity location.

The paper doesn't mention the socio-demographic variables that were found most influencing, but it remains clear that the list above contains mobility characteristics that are worth to be investigated on their correlation with household characteristics.

Renni Anggraine (2014) investigated the doing or not doing of an activity of two headed households. The activities of table3 were investigated.

No	Activity	Personal (P) or Household (HH) Level	Scope of Activities
1	Work	P	Full-time and part-time
2	Business	P	Work-related
3	Other	P	Other mandatory activity (school, etc)
4	Bring/get person	HH	Drop-off/pick-up children/spouse to a certain location
5	Shop-1-store	HH	Shopping 1 store
6	Shop-n-store	HH	Shopping n stores
7	Service-related	HH	Renting movie, getting (fast) food, institutional purposes (bank, post office, etc)
8	Social-independent	P	Visiting friends, relatives , etc
	Social-joint	HH	Same joint
9	Leisure-independent	P	Sports, café/bar, eating out, movie, museum, library
	Leisure-joint	HH	Same joint
10	Touring-independent	P	Making a tour by car, bike, or foot (e.g., letting out the dog, etc)
	Touring-joint	HH	

Table 3: modeled activities (scan from source, Renni Anggraine, 2010)

The variables that were found to be most important on doing the activity or not, were in order of importance (Renni Anggraine, 2010):

- Household activity type;
- Day of the week;
- Number of bring/get activities already included in the schedule;
- Number of children;
- Having a driving license;
- Number of employees in daily good sector within 3.1 km from the house.

Some of these factors are rather personal then household ones, but these two may overlap.

3.3 Environmental characteristics and mobility outcomes

The research about the New Danish Model (Jeppe Rich, 2010) was already discussed in section 3.1. Besides personal characteristics, their population synthesizer uses also the effect of Labor Market Areas to predict mobility outcomes. This was repeated here because it is an environmental factor and not a personal one. It was already explained previously.

There was stated that factors such as car ownership, employment and residential situation influence the activity behavior of people (and hence should be predicted to evaluate land use policies). On the other hand, supply of traffic accommodations have their results on decisions made by people as well (Shiftan, 2008).

"Models should include various longer-term individual and household lifestyle decisions, such as residential location, employment and workplace, auto ownership, and other potentially long-term activity commitments and long-term travel-related decisions, such as transit and parking arrangements."

Further, the paper recaptures this using figure 9.

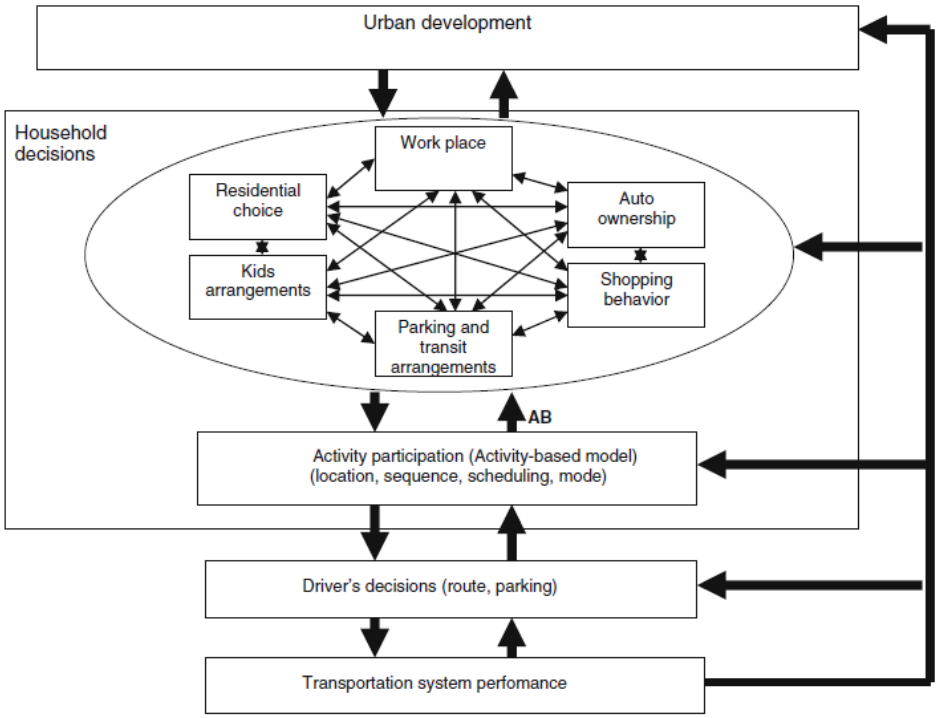

Figure 9: decision framework for performing activities

The relationships revealed can be read directly from the figure. For the main research, these remain interesting (a mobility variable related to a socio demographic variable)

- Residential choice along with work place and car ownership (interpreted as distance to workplace);
- Shopping behavior and car ownership;
- Residential situation and car ownership.

Further relationships remain not concrete enough in the source to see them as an indication of correlation for the main research.

4. Research Question

Following the aim of the research, the main question is:

"What personal/household variables (such as age, gender, income level, car ownership, …) have statistically significant correlation with characteristics in travel behavior (such as route choice, destination choice, travel frequencies, modal choice, …), based on social mobility databases?"

This research question remains quite vague; will there be accents on age, gender, income, … on the personal side and trip length, car occupation, trip frequency, … on the mobility characteristic side? In order to plan the research, some sub questions should be made (e.g. "What is the correlation between number of children in a household and trip frequency?").

In order to solve this, a database will be explored (see further). Decision trees will be made out of the database, giving probabilities one entity in the database may have at each attribute (e.g. number of children). This is further explained in chapter 5.

The literature review (chapter 3) is inspiration to define which variables will be researched. In fact, an initiation is given by literature (chapter 3). Afterwards, via decision trees this initiation is further explored and checked and eventually expanded (chapter 7). A possible correlation will then be found. This correlation is explored by building models (chapter 8). The conclusions are in chapter 9.

5. Methodology: used software and database

To build and analyze a decision tree, the program "Weka" will be used. It is freely available from the internet.

To research correlation, the program SAS will be used.

This research is not meant to become a guide on using software programs. However, to reach results, there must be used software because the available data is too big to handle manually. There was proposed to use the package WEKA and use it to build decision trees using the J48 algorithm. Therefore, this chapter describes some of its features to make a conclusion whether it is suitable to use for this research.

5.1 Properties of WEKA

WEKA is open source software developed by the University of Waikato by the programming language Java for data mining. This means it can work data to uncover patrons, correlations, fractions … that would otherwise remain invisible. Common techniques are association (dependency modelling), clustering (discovering groups of similar data), classification (classify data according to a class), regression (function building) and summarization (visualization, reporting …). WEKA provides some algorithms to build decision trees. This concept is explained below (Dr. Neeraj Bhargava, 2013). To use weka with as many data as possible (the database used in this research), the user must convert the data to ;arff extension.

5.2 Decision trees

Decision trees are data mining tools representing the probability of an instance of the data belonging to a combination of classes. For example, a decision tree could split a database into the category of sex and next, splitting each gender class into different age groups, assigning each instance of a particular gender class a probability of being in this group. Further attributes can be added and finally, a certain outcome will be visualized (e.g. probability of females in some age group on having a trip in a specific category).

Decision trees consist of root nodes (starting nodes, “gender” in the example), leaf nodes (the probabilitys the example) and internal nodes (test nodes in the tree, in the example “male”, “female” or the “age groups”).

The advantages of this tool are:

- Easy to understand;
- Many forms of data can be analyzed (nominal, ordinal, textual, …);
- Can deal with missing values;
- A tree can be pruned (see further);
- Most powerful approach in data mining and knowledge discovery.

In this research, the first step to answer the main research question is to find out which correlations between household/personal variables are worth being investigated, since it is impossible to unravel all mathematical relationships. First, the personal or household attributes influencing a mobility outcome needs to be identified. This can be done by the concept of pruning; a technique to keep a decision tree small and hence, eliminating unimportant internal nodes.

The classification attribute of each decision tree is a mobility characteristic (e.g. trip length distribution). For clarification, a possible example is given in figure 3 (with random chosen numbers).

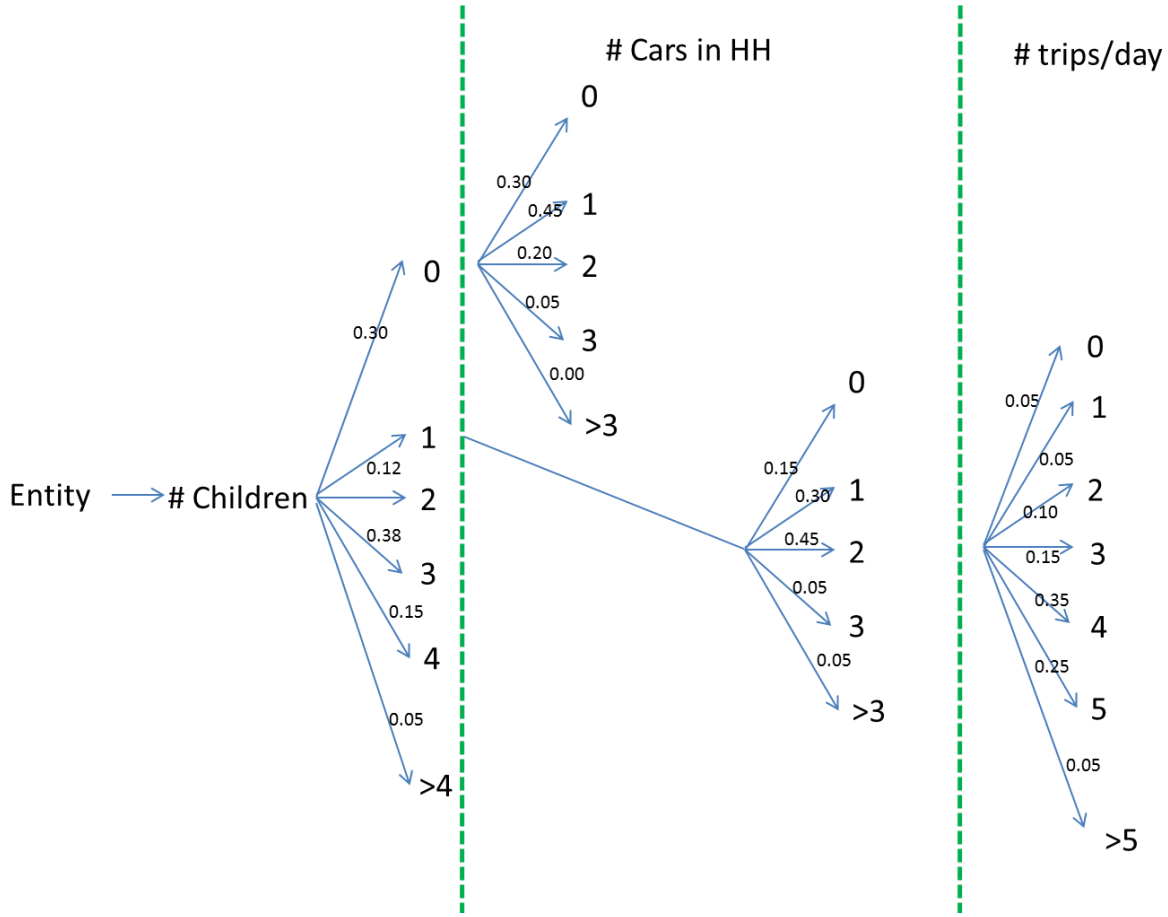

Figure 10: decision tree, variables #children, #cars, #trips/day (dependent variable)

To keep figure 10 clear, the decision tree only has the explanatory variables “number of children”, number of cars” and the dependent variable “trip frequency (# trips/day)”. For the same reason, it is only displayed for the outcome 1 of the variable number of children and outcome 2 of the variable number of cars. The result should be a distribution of the dependent variable for each possible combination of outcomes of the explanatory variables. This will justify the research questions, since there exist computer programs that reveal which variables cause a statistical difference in the distribution of the final outcome variable (e.g. when the distribution of an input variable, let’s say number of cars, does not differ among the possible outcomes in the previous step, i.e. there is the same distribution for each outcome of the previous variable, this previous variable will not be investigated as there is no correlation expected between the input variable (household characteristic) and the output variable (mobility characteristic). To clarify this, see figure 11.

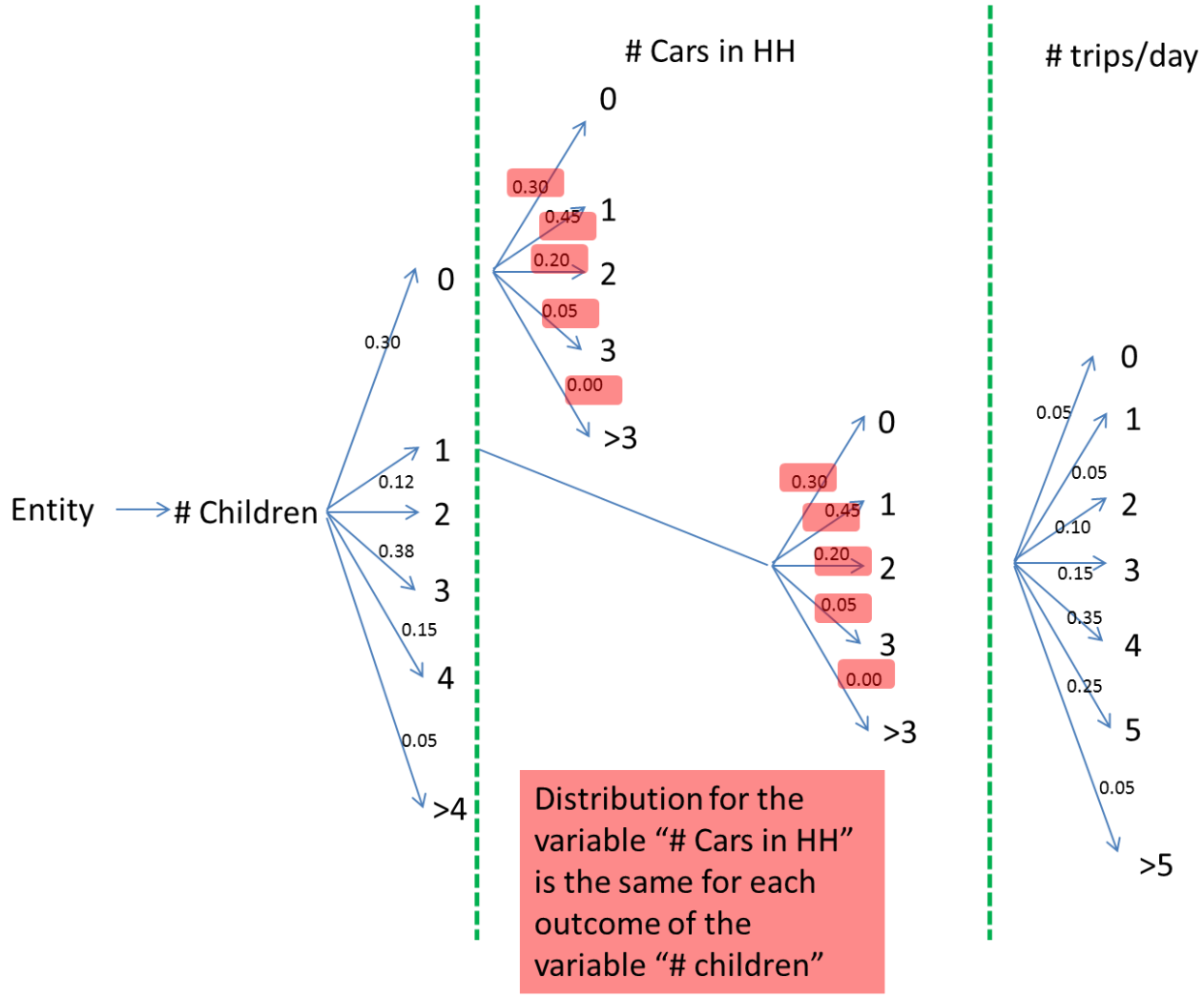

Figure 11: identical distributions for an explanatory variable

In the hypothetical example of figure 11, each distribution for the variable "# cars in HH" is identical among every outcome of the previous variable in the decision tree. This should make the conclusion towards research questions that the correlation between number of children and trip frequency will not be investigated.

Trees can be univariate and multivariate. Univariate trees do tests for a record on one attribute each, while multivariate trees do a combination of tests on records (e.g. the sum of income and age must transcend a certain degree). Trees in this research will be univariate. Firstly, the attributes that are likely to correlate are eliminated (e.g. distance to work and time to work) and besides, when there is suspicion of correlation between two attributes, this is researched and argued in later steps when the regression was made.

To build a small and efficient tree, the splitting should be based on the highest gain of information. Entropy is a measure (many times measured in bits) that shows the disorder in data. It is also called measurement of uncertainty in any random variable. It is the probability that a certain state of an

instance in the database occurs (Dr. Neeraj Bhargava, 2013). In the source for this section is an application of this algorithm that provides more insight in the theory. The information gain is calculated as:

$$Gain = entropy\_before\_split - entropy\_after\_split$$

The entropy before split is the entropy in the complete database, before it was tested on a certain condition of an attribute (similar for the entropy after). This can be calculated (for an example where the database is splitted on one attribute, it goes similar when there are more possible states):

$$Entropy\_before\_split = -a/t*log(a/t) - (b/t)*log(b/t)$$

Where:

- a = number of instances in complete dataset with classification attribute value = a;
- b = number of instances in complete dataset with classification attribute value = b;
- t = total instances.

After the split, the instances having a particular value on the attribute will be distributed in sub datasets, having each a specific value for ANOTHER argument. The entropy for these new groups is calculated with these formulas:

$$Entropy\_group\_1 = -a/t*log(a/t)-b/t*log(b/t)$$

Where:

- a = number of instances in group after first split, having attribute value a on the SECOND attribute
- b = number of instances in group after first split, having attribute value b on the SECOND attribute
- t = all instances remaining in group after FIRST split

The same formula must be applied for the instances where the FIRST attribute value is a different one than group 1 has.

Finally, in order to calculate the entropy gain, the total entropy for after the split must be calculated:

$$Entropy\_after\_split = a/t*entropy\_group1 + b/t*entropy\_group2$$

Where:

- a = total instances having attribute value a for first argument
- b = total instances having attribute value b for first argument
- t = total instances before any split

Applying this technique, a computer program builds the smallest possible tree, keeping as much information as possible. The important attributes get filtered among the ones that should gain more entropy; i.e. the groups belonging to one combination of attribute values become larger, automatically pointing out the most relevant attributes for one final outcome attribute (in this research, it could be for instance trip length category, while the attributes could be e.g. gender, age and income). The algorithm above makes sure that the top nodes of a decision tree provide the biggest information gain and have the most explanatory value for the investigated mobility characteristic.

After building the decision tree, leaves having not enough instances, are skipped. This is called pruning.

This text points out that the J48 (name for the C4.5 algorithm in WEKA) algorithm builds a small but accurate univariate tree, eliminating outliers (pruning) and selecting attributes on their accuracy. This is done because the algorithm subdivides the large group of data in groups having similar values on attributes so that the information in the dataset gets bigger. The most important attributes (i.e. the ones that provide the most information gain), remain in the final tree. This is exact what the research needs to develop its research questions. The research is not handed out by computer scientists, so the use of open source software permits a lot of online information available to the researcher. Hereby, the use of the algorithm and the program WEKA are considered to be justified.

5.3 Choice of the data set

For this research, three types of data sets have been explored for their amount of data and the quality of the research. It becomes clear that the database of the NHTS is the largest database and it is the one which has done the most effort for obtaining reliable data (payed respondents, double checking, trained interviewers …). Also, it is freely available. This makes it a good data set to derive significant statistical correlation between mobility data and personal data.

5.4 Description of database

In this paragraph, a description of the available files is given. Each file is referenced to a table in the appendix listing the number of attributes and a short description. The files discussed are limited to files giving results, not files to evaluate the database (such as weighting households in function of their participation degree).

5.4.1 Household file

The household file describes the participating households. Each household was assigned one identical number, which couples other mobility characteristics in the trip and tour files (see further). The attributes are in table A (see appendix).

5.4.2 Persons and their travel day

This file contains some mobility characteristics joined with household characteristics. All attributes from table 1 are copied, besides VARSTRAT, WTHHFIN, HHRELATD, RESP_CNT, SCRESP, WRKCOUNT and CNTTDHH. The variables from table B (see appendix) were added. The content of this file is similar to the results derived from the travel diaries from the already discussed OVGs.

5.4.3 Person file

This file contains some attributes about the respondents and personal, overall mobility characteristics (not related to one tour or one trip) such as number of bike trips last week, number of public transport trips last month … . The attributes are listed in table C.

5.4.4 Vehicle file

This file joins each participating household with a certain vehicle type with attributes as emission, fuel type … . It would be possible to make predictions about these attributes, but that is not the main goal of this research as it is to predict mobility variables. It is an opportunity for further research.

5.4.5 Tour file

A tour in mobility depicts a chain of trips, starting and ending in the same point, e.g. leaving the house, dropping children at school, continue to work, leave work to a supermarket, pick up the children and get back home (National Hosehold Travel Survey, 2011). This file couples each tour with the unique household and person ID of the participator who reported it and hence, makes it possible to join tour and household/person variables. The attributes are in table D (see appendix).

5.4.6 Chain trip file

As mentioned in 5.4.5, each tour consists of several trips. This file describes each segment of the tour with the attributes from table E(see appendix).

6 Preparation of the database

This chapter describes how the files of the NHTS are prepared to be used by this research for analysis. It doesn't go into detail on preparation for specific types of analysis, which will be discussed at the respective chapters, but for preparation to a starting point for preparation for detailed analysis.

6.1 Joining process

The database was explored and there was chosen to join the household file, travel day file, person file, chain trip file and tour file. Although the information of trip sequence is lost, what remains is a very big file (380 000 records) describing really any socio demographic variable in the database and any trip or tour characteristic in the database. This gives the research the opportunity to build very good models (since the large amount of data). Afterwards this big file was worked to fit to specific purposes.

The personal file will be linked to the household file and afterwards, this will be linked to the tour file. This is possible if there is a primary key created. Every record in the personal trip file can be joined to the tour file because of two columns describing household id and person id, where person id is meant to be the unique person number in that household. This way of identification was used in the tour file too. When combining these two columns, each person has a specific id and the tour file can be coupled with the personal trip file. Afterwards, it is joined with the household characteristics. This should deliver a large database, having records indicating mobility variables with many explaining variables.

There is a problem with the person file: when importing the dataset, the variable WRKTIME, which stands for “usual arrival time at work”, is only filled in once per person id. For the other records, it’s marked with a -1, indicating this is a missing value. However, this permits the research to simply skip these records, keeping one record per person in the file, as in this research part correlation between personal and household variables and the tour type variable are sought.

Afterwards, the variable WRKTIME was skipped completely due to technical issues. There was not expected that keeping it was worth the effort of data formatting.

The joining process is in figure 12.

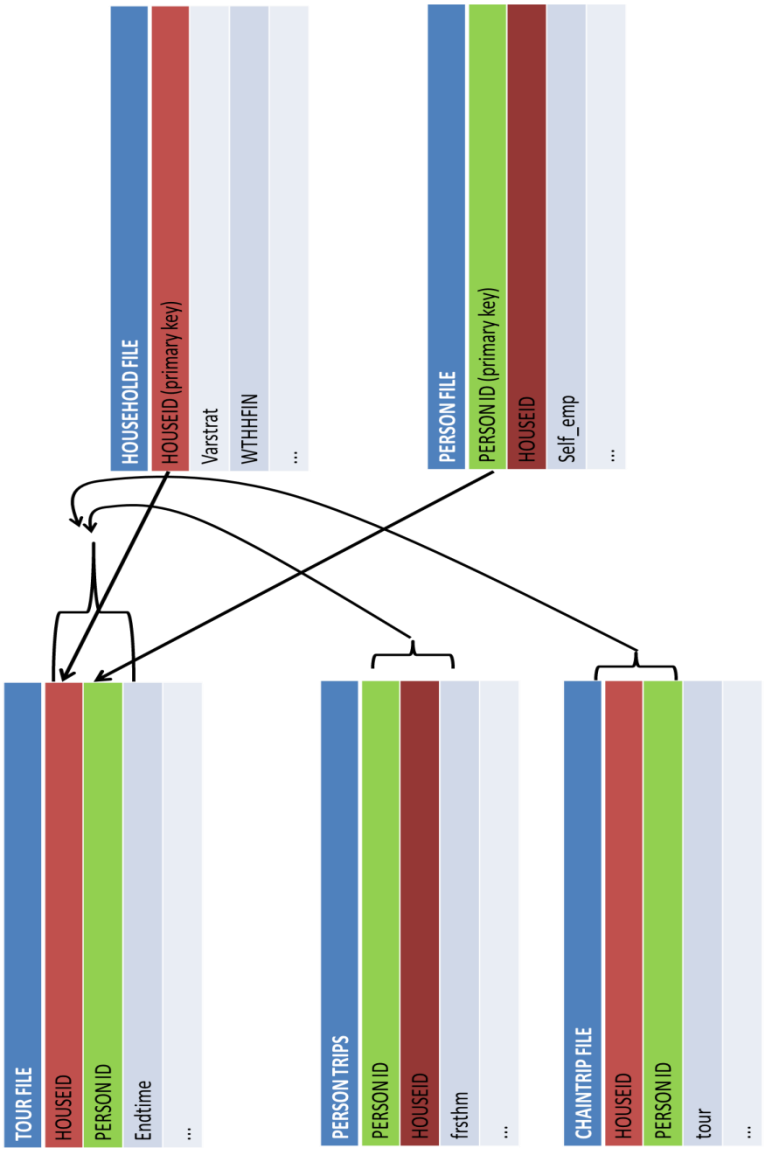

Figure 12: joining process for NHTS database

6.2 Omitted type of variables

For this, the goal of the research is revisited. This is to build universal functions indicating the relationship between personal characteristics and mobility characteristics in order that they don't have to be estimated for each transportation model anymore. Instead, the mobility characteristic can be deduced from variables directly available (from demographic databases).

Therefore, these kind of variable will be omitted:

- Variables used for internal consistency in the National Household Travel Survey Database (e.g. household weight, filled in completely, …) since this research only uses results and does not try to evaluate or improve the database;
- Variables providing information of consequences of travel behavior (e.g. emission per traveled mile per vehicle type) as it is a later stage of policy;
- Attributes describing respondent's locations. However location will be present in the research, this will be measured by transferable attributes (city size, rural area, …) instead of direct positions. Otherwise, it is hard to transfer the data to non US areas;
- Attributes providing information that needs an inquiry itself to be captured (e.g. respondents opinions). If this research has to be done, it abolishes the research goal;

Furthermore, variables functioning as a key between files (HOUSEID, PERSONID, …) are not omitted, but only used to produce analysis files and hence not in the final analysis.

7. Construction of decision trees from the NHTS database

This chapter describes the building of decision trees: why are they built and what can be learnt from them for the later research? Note that these decision trees are not built to quantify the effects, but just to indicate **WHICH** variables are important to predict a mobility outcome (and not **HOW** important). This is because the technical side is quite complex and it is difficult to change code in this dataset. In later stages (building models), these variables are all adjusted and their effects are quantified.

7.1 Household and personal characteristics and tour type

Following the literature review, there seems to be a correlation between personal and household characteristics and the type of activity and hence, type of tour (as presented in the NHTS).

Tour type stands for a tour originating from "home", "work", or "other" and arriving at one of these categories (9 combinations possible) (McGuckin, 2004).

It is also important to note that a tour is not necessarily seen as a tour! In the NHTS database, every trip is seen as a part of a tour and that trip is assigned the tour type of the tour it is part of. This was done by making one big file of the parts of the database.

The resulting decision tree is in figure 13:

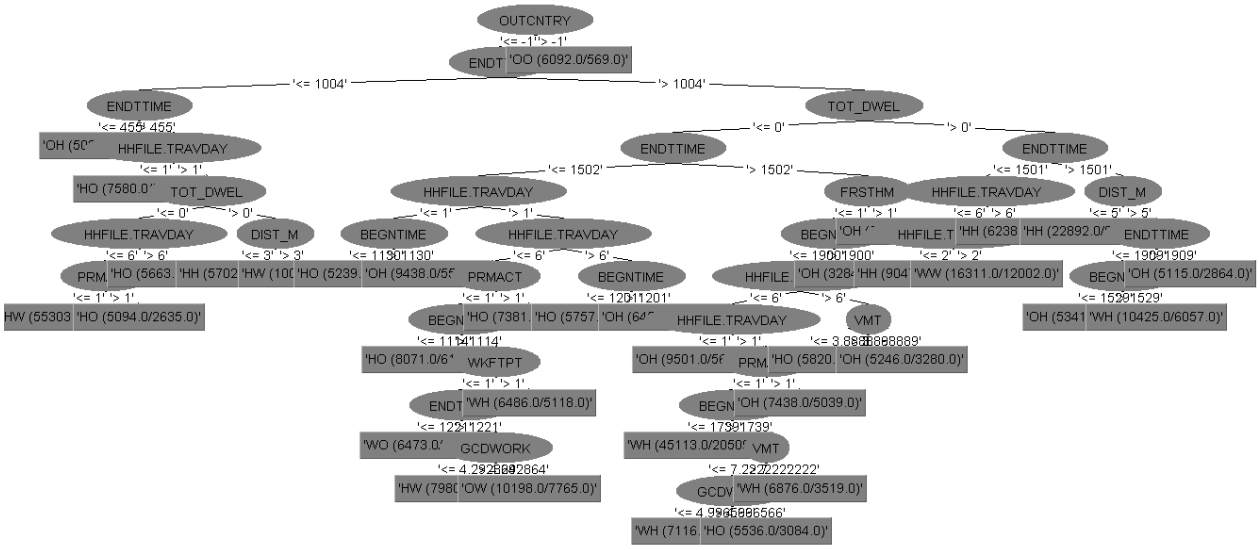

Figure 13: decision tree tourtype

Due to graphical limitations, it may be hard to read the important variables. However, explanatory variables for the mobility characteristic of tour type are:

- If the respondent left the country;
    - This is logic, because this is some tour type on its own. Therefore, it is not taken into account any further.
- The end time of the tour;
- The day of the week;
- The begin time of the tour;
- Distance traveled;
- Distance to work;
- The primary activity of the last week;
- If the respondent started his travel day at home;
- The time spent on the destination.
    - This was skipped as well, because it has a direct link with the tour type and therefore, it must be an explaining variable.

These variables are found to be important following the decision tree. After testing this with the literature review, the variables distance to work and distance traveled are one of the variables that remain. Jeppe Rich (2010) stated that labor market areas have a significant combined effect with age on mobility as a whole, but in this paper, it was used for different models for different times. It is thinkable that different times can result in different tour types. Davy Janssens (2005) stated in a research that trip length is an important factor to include in models. The variable “day of the week” is also kept. Renni Anggraine showed in 2010 that this factor was important of doing an activity or not, researching different types of activities. The research of Linda Nijland (2014) confirms this, especially for shopping

activities in the weekend. Linda Nijland also showed that the parameter time (availability and time passed since last performing the activity) is an important one in doing different activity types. Therefore, the variables end and begin time of the tour and the primary activity performed last week are not rejected in this stage. This was also confirmed by Arentze and Timmermans (2002).

7.2 Influence of personal and household characteristics on distance traveled

The research travel behavior for Flanders (see literature review) states a correlation between sex and traveled distance, mode choice and trip motive (business vs non business). Jeppe Rich (2010) suggests that the effect of gender is combined with age. Trip motive was already discussed above and will be skipped in this paragraph. Davy Janssens (2005) suggests that mobility outcomes as trip duration and length (depending upon each other) are influenced by household and personal characteristics (the specific household and personal characteristics were not mentioned, because the aim of that research was to compare two analyzing instruments, i.e. Bayesian networks and decision trees).

This paragraph researches the relevant personal and household variables influencing traveled distance and mode choice. Based on literature, there should be a correlation with gender and age. It is thinkable that other characteristics are also important.

In the joined file, the attribute tot_mils indicates the total tour length. Variables directly depending on tour lenght (e.g. the length of the longest segment) will be omitted, done in a same way as in the previous section, along with variables for internal database consistency and identifying variables (they have no explaining value and were only used to join separate files). Finally, geographic variables were deleted because their meaning is pictured by other (more general) variables (e.g. a variable describing the state a household lives in gets described by other variables such as population density, rural/urban environment …) Here is stressed that particularly variables describing the duration of a tour are deleted as well!

In this file, relevant household and personal characteristics were linked with the total miles a tour contained.

Max: 5600 miles

Min: 0 miles

Average: 14.5 miles

Median: 7 miles

Based on these statistics, tours that contained more than 200 miles were assigned to a class of their own, since there are only 1905 records having it (that's 0.5% of the total reported tours). This number was chosen because of the great difference between the minimum and the maximum. It would be inefficient to make classes from 0 to 5600 miles, when the average is 14.5 and the median 7. When randomly trying to impose 20 miles as the start for the remaining class, it seemed that the proportion of records was still quite high. The start of the remaining class was experimentally altered until its

proportion would become quite small, which was the case at the 200 miles level. With a level of 100, it remained 1%. There will be made two classes to deal with the high distance tours (tours having more than 100 miles, which is 1% of the database). A class “from 100 to 200” and a class for “200+”.

Under 100 miles, there will be made groups in steps of 10 (table 4).

GROUP	DISTANCE LEVELS (miles)
A	0 to 10
B	11 to 20
C	21 to 30
D	31 to 40
E	41 to 50
F	51 to 60
G	61 to 70
H	71 to 80
I	81 to 90
J	91 to 100
K	101 to 200
L	more than 200

Table 4: distance classes

The level assigning was done by this excel function:

```
=IF(AND(A2>=0;A2<=10);"A";IF(AND(A2>10;A2<=20);"B";IF(AND(A2>20;A2<=30);"C";IF(AND(A2>30;A2<=40);"D";IF(AND(A2>40;A2<=50);"E";IF(AND(A2>50;A2<=60);"F";IF(AND(A2>60;A2<=70);"G";IF(AND(A2>70;A2<=80);"A";IF(AND(A2>80;A2<=90);"I";IF(AND(A2>90;A2<=100);"J";IF(AND(A2>100;A2<=200);"K";”L”)))))))))))
```

The distribution is as follows (figure 14):

number	category
248587	A
71881	B
28651	C
12717	D
6363	E
3508	F
2155	G
884	I
724	J
2808	K

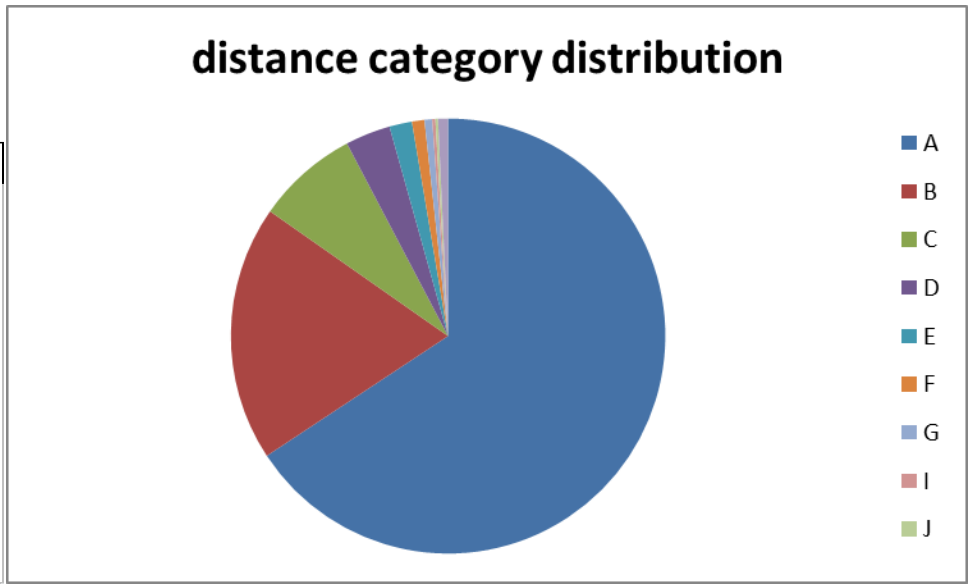

Figure 14: distance distribution

It is clear that, when the tree gets pruned, only records with distance category "A" remain. To get rid of this, distance class A was divided in 3 subclasses. The categories now become (table 5):

GROUP	DISTANCE LEVELS (miles)
A1	0 to 3.3
A2	3.4 to 6.3
A3	6.4 to 10
B	11 to 20
C	21 to 30
D	31 to 40
E	41 to 50
F	51 to 60
G	61 to 70
H	71 to 80
I	81 to 90
J	91 to 100
L	101 to 200
M	more than 200

Table 5: adapted distance classes

The excel function now becomes:

```
=IF(AND(A2>=0;A2<=3.39);"A1";IF(AND(A2>3.39;A2<=6.39);"A2";IF(AND(A2>6.39;
A2<=10);"A3";IF(AND(A2>10;A2<=20);"B";IF(AND(A2>20;A2<=30);"C";IF(AND(A2>30;A2<=40);"D";IF(AND
(A2>40;A2<=50);"E";IF(AND(A2>50;A2<=60);"F";IF(AND(A2>60;A2<=70);"G";IF(AND(A2>70;A2<=80);"H";I
F(AND(A2>80;A2<=90);"I";IF(AND(A2>90;A2<=100);"J";"K"))))))))))
```

The distribution of distance categories now becomes as in figure15. It is clear that it is more equally distributed and hence, the algorithms will be able to determine explaining variables from the file.

number	category
116021	A1
71402	A2
59806	A3
71881	B
28651	C
12717	D
6363	E
3508	F
2155	G
1358	H
884	I
724	J
4713	K

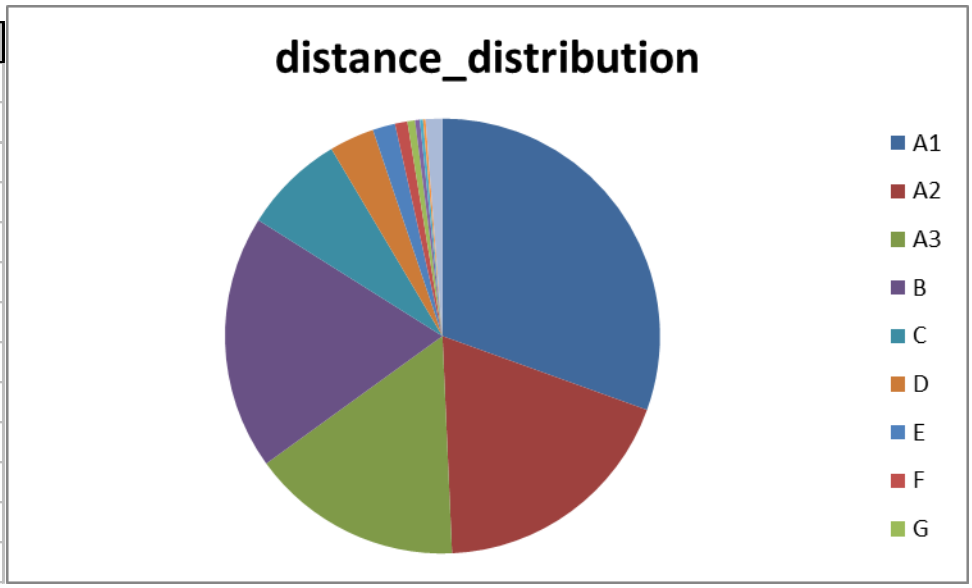

Figure 15: modified distance distribution

After conversion to .arff extension, it was ready to be loaded in WEKA. The resulting decision tree is in appendix 2.

This tree was pruned with a minimum of 1000 instances per leaf (to keep the analysis possible by hand). It is clear that these variables are important:

- distance to work (DISTTOWK);
- mode for longest segment (MODE_D);
- tour type;
- population/work density (HTPPOPDN/ HTEEMPDN);
- number of stops (STOPS);
- interstate used (USEINTST);
- begintime (BEGNTIME);
- urban or rural area (URBRUR);
- number of travel days (CNTTDR, especially combined with number of stops and use of interstate highway system);
- great circle distance in miles between home and work (GCDWORK);
- percentage of renter occupied housing (HTHTNRNT;

- part of a tour or not (TOUR_FLG).

These are intuitive results, but may have internal correlation. Also, there are some mobility characteristics explaining the mobility characteristic “distance category”: “mode for longest segment” and “tour type”. The analysis of the variable “tour type” (chapter 7.1) also showed a correlation between tour type and distance to work and traveled distance, insinuating a correlation between traveled distance and distance to work. To complete the analysis, a new tree will be built with the file used before, but eliminating the variables “mode for longest segment” and “tour type”, in order to see purely personal or household variables influencing the traveled distance. To be sure, mobility variable “mode for longest time segment” will be skipped too. When this is done, the variables “end time” and “travel day” become more important. As a summary, these variables influence traveled distance:

- Distance to work
    - Cf. great circle distance between home and work
- Area description
    - Population density
    - Percentage of population at work
    - Percentage renter occupied
    - Urban or rural area
- Time stamp (not time traveled, this was skipped on purpose, see earlier)
    - Number of days traveled
    - Begin time
    - End time
    - Travel day (of week)
- Characteristics describing the tour on its own
    - Part of a tour or an independent replacement
    - Interstate highway usage
    - Number of stops

This last category is seen as rather a consequence of the traveled distance than variables explaining it and will not be discussed further.

The importance of the variable “distance to work” is not surprising, as 43.5% of the trip motives arrive or originate at work (165391/380183). The frequencies of the tour types are in table 6. The correlation between distance to work and great circle distance to work was checked (see further).

Number of observations	TOURTYPE
36976	HH
68805	HO
67628	HW
75208	OH
33803	OO
10809	OW

Number of observations	TOURTYPE
59939	WH
17973	WO
9042	WW

Table 6: distribution of travel motives

The variables that describe the environment of a respondent, could be correlated, as the variables population density, population at work and percentage renter occupied cover the same area and hence, the same people. It is thinkable that these values differ for urban or rural areas and hence, a correlation between these variables was investigated as well. This will be discussed further (HRSA, how is rural defined?, 2015).

7.3 Household and personal characteristics and mode choice

As stated, the research travel behavior for Flanders (D. Janssens, 2014) suggests an influence of sex on mode choice. Zhizu Zhanga (2013) and Ibrahim Seikh A.K. (2007) also found influence of social variables influencing mode choice.

The NHTS database provides the possible modes from table 7 (second column). It is given for the longest segment of the tour. For this research, the range of possibilities is too complicated; the NHTS also provides specific vehicle information (gas usage, vehicle age …) and the extensive mode description may be rather important when also that information is used. Therefore, it is simplified as in the third column of table7.

Code	Meaning NHTS	Simplification
1	Car	Privately Owned Vehicle (POV)
2	Van	POV
3	SUV	POV
4	Pickup truck	POV
19	Taxicab	Taxi
22	Bicycle	Bicycle
23	Walk	Walk
5	Other truck	Truck
7	Motorcycle	Motorcycle
9	Local public bus	Public Transport (PT)
10	Commuter bus	Public Transport
11	School bus	Public Transport
12	Charter/tour bus	Public Transport
13	City to city bus	Public Transport
14	Shuttle bus	Public Transport
15	Amtrak/inter city train	Public Transport
16	Commuter train	Public Transport
17	Subway/elevated train	Public Transport
18	Street car/trolley	Public Transport
8	Light electric veh (golf cart)	Others
24	Special transit-people w/disabilities	Others
97	Other	Others

Table 7: possible modes

The resulting decision (figure16) tree shows that almost nothing but traveled distance (TOT_MILS) and (which is likely to be correlated) traveled time (TOT_CMIN) matters. This is confirmed by Ibrahim Sheikh (2007) in his research “Mode Choice Model For Vulnerable Road users in Malaysia”. However, this tree suggests that is very likely that trip distance and time are one of the most important factors for mode choice. However, Zhizu Zhanga showed in 2013 that personal characteristics also have their influence. To find more influencing variables, a second tree was built, leaving out the variables TOT_MILS and TOT_CMIN. There were also less instances per leaf needed for this tree (1500 instead of 3000). This one is in figure 17.

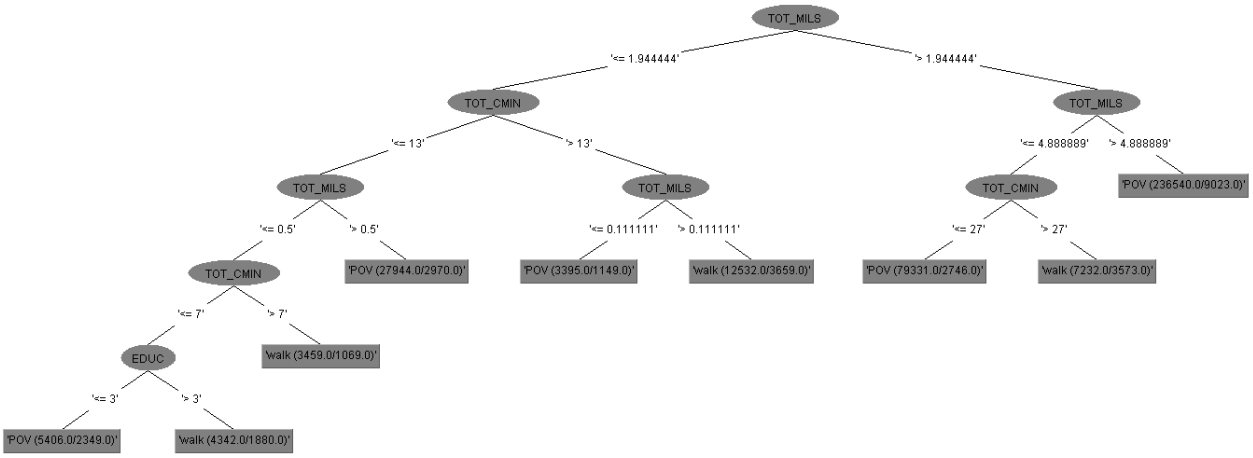

Figure 16: decision tree with distance and time variables mode choice

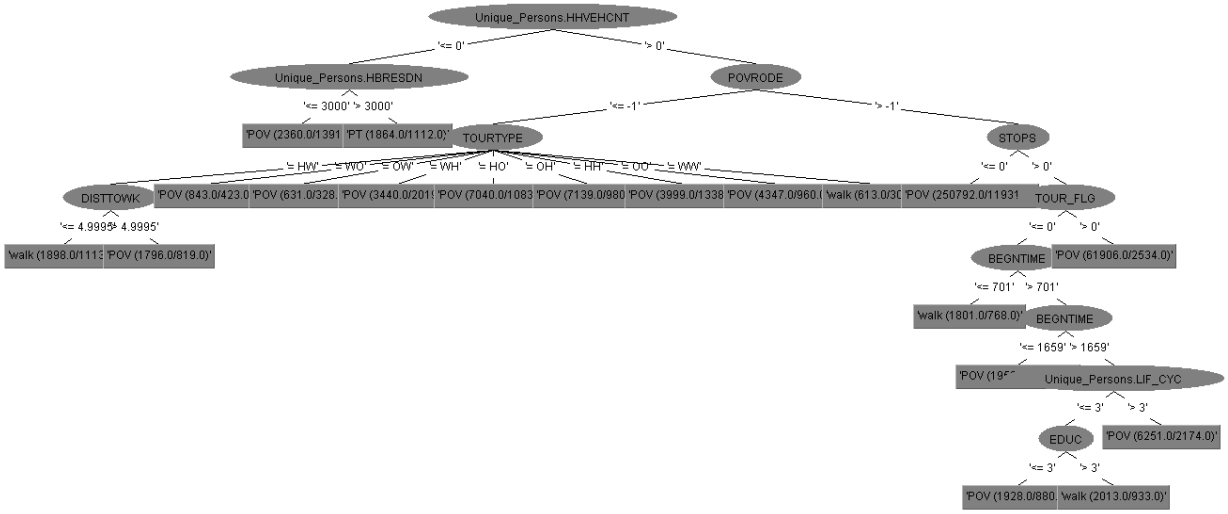

Figure17: decision tree without distance and time variables mode choice

At first, the tree was split on the variable vehicle count. This is not surprising, since in many cases one first needs to own a car before using it. The variable HBRESDN, describing the density of the built environment, splits the group into people using private vehicles and people using public transport (higher density than own vehicle users).

Tour type is not very influencing, only when it is a trip arriving at workplace. Then it is clear that the variable distance to work has its influence. Of course, it must be that this variable also influences mode choices for trips originating at work (the coming back trip).

The variable POVRODE is about the number of people in the car for the last week (when the diary was filled in). However, this variable is only split between the values greater or smaller than -1, this being the code for not knowing or not having used a car last week. Hence, it may be that people who use a car with more people (a higher occupancy rate), are more influenced by the type of tour they make to choose their mode. Specifically, the tour being really a tour with stops (TOUR_FLAG>0) or just a single ride is important. It seems that time of day has a role in mode choice for single rides (BEGNTIME). Also, the type of family (number of adults, having or not having children …) has its effect.

A last thing, is that the variable EDUC appears in **both** trees; a typical personal variable describing the educational level of the respondent.

It appears that one of the exploring researches (Mode Choice Model For Vulnerable Motorcyclists in Malaysia) for this thesis mentioned more personal variables for mode choice, not coming back in the analysis of the NHTS database. There was mentioned that income has an effect (however this effect comes partially back in variables as life cycle). It may be that the USA has a higher income level and that it therefore has less influence ($10,829 in Malaysia, 54,629 in the USA, (World Bank, GDP Per capita)). Also, gender has here not a proven effect. It may be that the aim of that research was for motorcyclists, a mode that possibly is more influenced by personal variables. However, the argumentation of using educational level can here be justified, as it can be an indication of income category.

7.4 Household and personal characteristics and activity duration

The NHTS database contains a variable called “TOT_DWEL2”, describing the time spent on a destination in minutes. Actually, it sums up all the stops in a tour, including short breaks. There is assumed that it is at least a very well indicator for activity duration and many times directly the time spent on destination. It was covered in the joined file discussed above. The New Danish National Model (Jeppe Rich, 2010) suggests age, gender and labor market area (the area one is able to commute within to his or her job) as explaining factors for trip duration (time on destination plus travel time). When exploring Bayesian networks versus decision trees by Davy Janssens (2005), there was found that activity duration was important to explore in combination with personal variables, however there was not really mentioned which variables. This thesis already defended modeling of activity type (see e.g. (Linda Nijland, 2014)), which is likely to be correlated with activity duration. Shiftan (2008) showed that land use has effects activity location, mode and **sequence and scheduling**. These last two variables include a time table and hence, activity duration. Arentze and Timmermans (2002) built the ALBATROSS model, which directly simulated activity duration based on long and short term socio demographic factors (marital status,

number of children, residential location, work place, work type, mode availability …). However, this was a simulation interfering with other, already made decisions and not a natural state of law. It is clear that there will be correlations between many socio demographic factors included in the NHTS database and the variable TOT_DWEL.

The variable TOT_DWEL2 is sometimes left blank. Respondents may not have filled this in or may not have known very well what was meant (what is an activity?). These problems arising with data capturing were described in the introduction. For this research, there was chosen to drop those records having no value for dwell time, simply because there is assumed that the travel day ends and there is not really an activity left of which the duration can be modeled. Further, the durations were divided in categories to apply the J48 algorithm as in table 8. This was done experimentally until each class would have enough observations to be able to determine the influencing variables. When modelling, the original continuous variable can be used.

Maximum: 1280 minutes (21,3 hours)

Minimum: 0 minutes

Time (minutes)	Category
]0-15[	A
[15-30[	B
[30-60[	C
[60-90[	D
[90-120[	E
[120-180[	F
[180-240[	G
[240-300[	H
[300-360[	I
[360-420[	J
[420-480[	K
[480-540[	L
[540-600]	M
>600	N

Table 8: activity duration classification

The distribution looks as in figure 18 and table 9. It is likely that the most important influencing variables will be found with this distribution.

act_duration_category	count
A	5221
B	6119
C	38311
D	29145
E	18013
F	22768
G	16424
H	13146
I	7696
J	5088
K	5458
L	12700
M	10823
N	7872

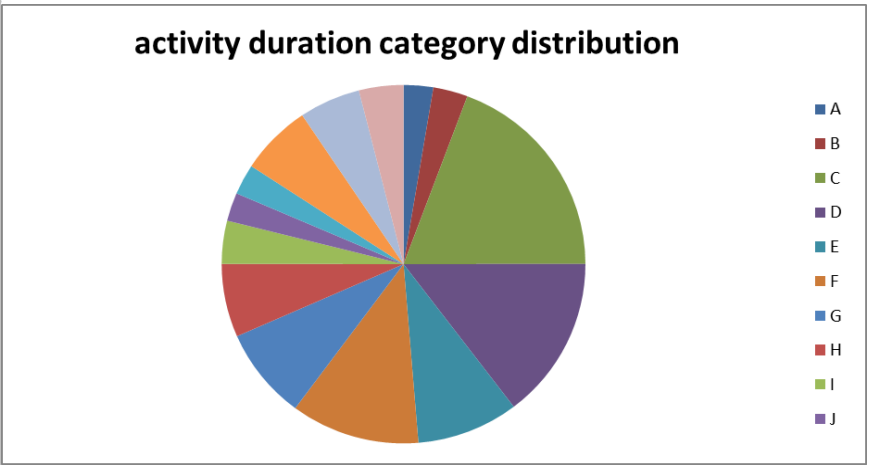

Figure18: distribution activity duration category

Table 9: distribution activity duration category

When a first run was applied, the variable tour type seemed to be the most important one. This is to be expected, but rather a mobility outcome than an explaining variable and the aim is to find natural laws to link behavior with person and household characteristics. Therefore, it is skipped. There can eventually be built a model for different tour types and the associated activity duration. It could be useful in activity based modelling, because the steps are taken sequentially and hence, tour type will be chosen in a previous step. The decision tree for activity duration is in figure 19.

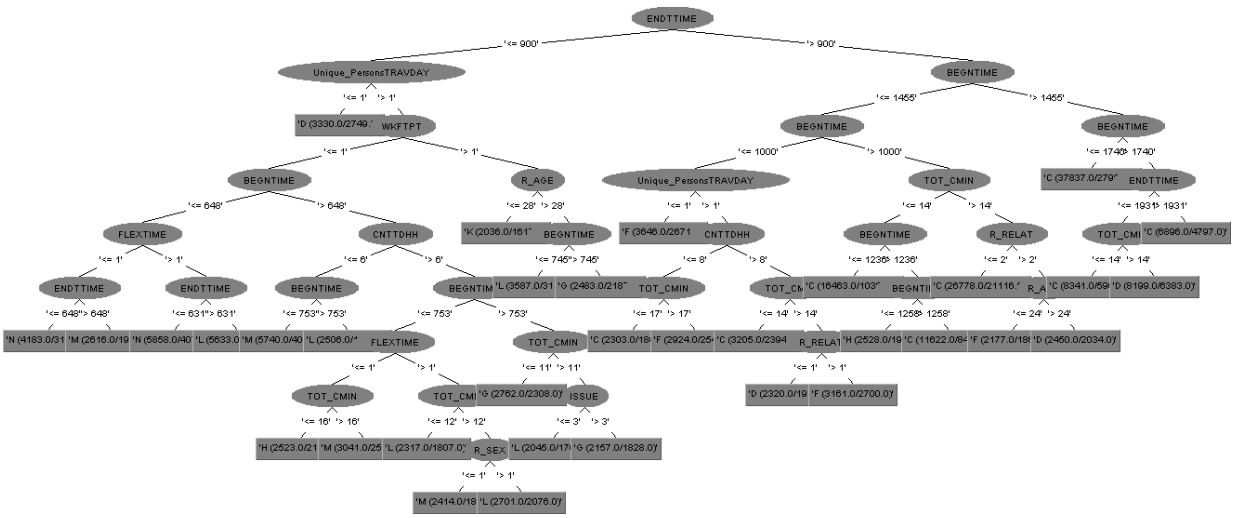

Figure 193: decision tree activity duration category

A first thing that stands out is that activity duration has an influence of time when a trip or tour (and so, the activity it is leading to) starts or ends. This was not a result of the literature review, but many models are built with a time component on its own (e.g. ALBATROSS by Arentze and Timmermans, 2002). There seems to be a correlation with the activity duration and tour length in minutes, with longer trips leading to longer activities. When researching travel time ratios (a concept that describes the percentage of time traveled of a time interval in which a certain activity was performed) this was confirmed; the travel time related to stay time can be considered to define someone's spatial reach. There was found that non - daily activities have a lower ratio; a smaller percentage of the interval in which the activity took place was spent on traveling (Vidacovic, 2000). This corresponds with the findings in this research, because a long trip will take a long activity, declining the percentage (not the absolute values).. Flexibility of respondents leads to different destination times, although it is not very important as it only appears once in the tree and in combination with the variable end time. However, part time workers and full time workers may have different results, as was also already stated by Linda Nijland (2014). The tree is split many times along the travel day being greater or smaller than one, meaning that there is a difference in duration classes for Sundays than for other days of the week (NHTS codebook).

Some typical household variables are important: relationship to household respondent (as in: husband or wife, child, parent …), age and sex.

Number of household trips has some effect on activity duration. This is not surprising, as there was already discussed that land use has its influence on activity behavior. Land use has its effect on number of trips (urban areas may have more but shorter trips (OVG Vlaanderen). People in urban and rural areas spend the same amount of time and money to traveling (JC, 1979).

Finally, there was splitted on the variable ISSUE. However, this variable indicates the transportation issues the respondents is aware of and is not some of the variables this research searches for modelling.

8 Models

Until now, the research has sought for indications that some household or personal characteristic influences mobility variables (Literature). Next, this was tested by building decision trees. The found correlations were discussed. Finally, the correlations have to be quantified by building models.

8.1 Tour Type Model

The model was constructed with 100496 instances due to missing values. The original database had 380174 instances. The parameter estimates are in table 10 to table 17. It is a generalized logit model and it predicts the probability of a tour type outcome in function of the variables mentioned before, having the tour type home work (HW) as a reference. This means that the probabilities are expressed as the difference with the probability of a HW tour type. Note that the abbreviations for different tour types are HW (home-work), HH (home-home), HO (home-others), OH (others-home), OO (others-others), OW (others-work), WH (work-home), WO (work-others) and WW (work-work). It is also important to note that a tour can be seen as a trip! In the NHTS database, every trip is seen as a part of a tour and that trip

is assigned the tour type of the tour it is part of. This was done by making one big file of the larger database.

As was stated, the variable day of week has a great influence. The first run was a generalized logit model with a dummy for each weekday with Wednesday as a reference, giving each weekday a dummy variable with Wednesday as a reference. Although many of the other explaining variables were significant, the other days of the week were not proven to have a different influence on tour type compared to Wednesday. There were found significant differences for Friday, Saturday and Sunday, which will be treated as the weekend days. Because weekday is a categorical variable, it resulted in a very big and complex model, each category (7-1) has a combination with each possible tour type (9-1), which results in 48 combinations. Because there seemed to be only difference between weekday and weekend day, the variable day of the week was split in just weekday or weekend day, reducing the 48 combinations to 6 combinations. The quality of the information delivered by the model is expected to be improved, but the quantity of the information is reduced.

The variable PRMACT (primary activity last week) also had an influence. This variable has 7 possible values (working, temporarily absent from work, looking for work, homemaker, going to school, retired, others). If dummies would be created for 6 possibilities, this results in 48 combinations (8*6, dummies for tour type multiplied by dummies for prmact). It makes the model very complex and it may be hard to find all the data to assign an agent to a primary activity in last week class. Therefore, as similar with the variables weekday, the categories for primary activity last week will be reduced to three: active agents (working, temporarily absent from work, homemaker), school going agents (going to school) and non – active agents (looking for work, retired, others). This results in two dummies (active agents taken as reference). This reduces the number of combinations to 16 (2*8).

The variable travel day started or not started at home is a dangerous one, because it may be in many cases directly define the tour type (HW, HO, HH). It is obvious that tours having the characteristic of not having started at home, will have a (very much) bigger probability to occur when the variable started at home is not true compared to tours of the type home – work, directly **not** having this characteristic and of course having the characteristic of having started at home. It was eliminated.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**HH**	1	-5.2624	0.0543	9385.1968	<.0001
Intercept	**HO**	1	-4.5603	0.0432	11122.0168	<.0001
Intercept	**OH**	1	-7.3907	0.0520	20227.9833	<.0001
Intercept	**OO**	1	-6.1392	0.0551	12400.5573	<.0001
Intercept	**OW**	1	-4.9604	0.0685	5238.5997	<.0001
Intercept	**WH**	1	-7.0578	0.0536	17309.2345	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**WO**	1	-5.3160	0.0607	7679.5007	<.0001
Intercept	**WW**	1	-4.5312	0.0763	3526.0440	<.0001

table 10: intercepts tour type model

The intercept estimates are all very significant (p-values <0.0001). This means that there is a clear difference in probability of the tour types compared to home-work tours if other explaining variables are not considered. This may be strange for the tour type work-home, because an agent who went to work, also is expected to have to come home again, which should result in a same probability as the reference home work. However, there may be intermediate stops, resulting in reported tours of another type.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
ENDTIME_IN_MIN	**HH**	1	0.00682	0.000185	1355.4429	<.0001
ENDTIME_IN_MIN	**HO**	1	0.00328	0.000143	528.3958	<.0001
ENDTIME_IN_MIN	**OH**	1	0.00464	0.000141	1076.6020	<.0001
ENDTIME_IN_MIN	**OO**	1	0.00412	0.000177	539.5916	<.0001
ENDTIME_IN_MIN	**OW**	1	0.00232	0.000208	124.0807	<.0001
ENDTIME_IN_MIN	**WH**	1	0.00502	0.000155	1043.5559	<.0001
ENDTIME_IN_MIN	**WO**	1	0.00301	0.000189	254.4724	<.0001
ENDTIME_IN_MIN	**WW**	1	0.00534	0.000244	478.4856	<.0001

Table 11: endtime parameters tourtype model

The difference of probability compared with HW trips caused by end time in minutes (to be calculated by minutes starting from 00:00) is significant for all tour types. It is very small, but the values can be very high. It is also positive for all possibilities. This means that every tour type has a bigger chance to occur than home - work tours and that this effect grows when the time advances. This is not surprising because tours starting at home and arriving at work are often early; giving other tours a larger probability later on the day. The biggest part of the US working population in 2015 works from 9AM to 5PM (Schawbell, 2011). In addition, USA workers work per person averagely 1789 hours per year (OECD, 2014). This means about 7 hours a day if weekend days are not considered. These hours will be more for people who do actually home - work trips, because non workers are also presented in this average. This makes it very thinkable that home - work tours occur earlier than other types. The parameters are relatively big for tours arriving at home. This corresponds with the reasoning above. Work – work tours (WW) also have a higher probability later on the day.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
weekend	**HH**	1	1.0185	0.0296	1182.4367	<.0001
weekend	**HO**	1	1.3871	0.0253	3011.3060	<.0001
weekend	**OH**	1	1.3316	0.0269	2445.7793	<.0001
weekend	**OO**	1	1.4619	0.0303	2330.9054	<.0001
weekend	**OW**	1	-0.0817	0.0479	2.9181	0.0876
weekend	**WH**	1	0.0638	0.0292	4.7595	0.0291
weekend	**WO**	1	*-0.0234*	0.0393	0.3559	0.5508
weekend	**WW**	1	-0.2209	0.0526	17.6628	<.0001

Table12: weekend parameters tourtype model

The parameter weekend describes the change of probability of tour type relative to the reference (HW tours) when a tour is in the weekend (here defined as Friday, Saturday or Sunday). Most estimates are very significant, but the types others-work, work-home, work-others are not proved to have a different probability of appearance in the weekend then in the weekdays **on a 95% significance level.** When the standards are lowered to a 90% significance level, there would be no problems for the tour types others-work and work-home.

Over all, there can be decided that all trips are proved to have a statistically different appearance in the weekend than during the week compared to the reference, besides for work-others relationships. However, this tour type remains vague; what is “others”? It could be anything. Since it is the only one that is really insignificant and because the tour type others-work is significant, there was decided to keep it in the model. After all, there is no reason to expect a difference in appearance for trips from “anywhere” to work in the weekend than from work to “anywhere”.

The signs of the parameters for work related tours are to be expected, but remain small in value. This is because it is not likely one does a work related tour in the weekend compared to the week. The small values are because the tour types are also compared to another work related tour type. An exception on the negative sign is the work related tours is the parameter of work home based tours. It may be because Friday is also seen as a weekend day. This was already explained before in this section.

Not work related tours have a bigger and positive parameter. It is an indication that people do other activities in the weekend. Other – other tours have the biggest parameter; it may be that people leave their home during the weekend.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
nonactive_people	**HH**	1	0.6893	0.0779	78.3947	<.0001
nonactive_people	**HO**	1	0.9573	0.0668	205.2917	<.0001
nonactive_people	**OH**	1	1.0373	0.0704	216.9795	<.0001
nonactive_people	**OO**	1	0.9139	0.0776	138.6264	<.0001
nonactive_people	**OW**	1	-0.1409 **0**	0.1436	0.9631	0.3264
nonactive_people	**WH**	1	0.1179 **0**	0.0822	2.0581	0.1514
nonactive_people	**WO**	1	-0.0325 **0**	0.1138	0.0813	0.7755
nonactive_people	**WW**	1	-0.6986 **0**	0.1890	13.6571	0.0002

table 13: non-active people parameters tour type model

Note that people looking for a job are also seen as non-active people and hence, there may be work related tours reported (e.g. a solicitation). Parameters for tour types HH, HO, OH and OO are very significant and have expected signs. Tours beginning or ending not at home or not at work have the biggest probability to be done by non - active people and that is statistically proven. Work related trips are not proven to occur more by non - active people than by the working population, except for WW tours.

Recap the variable where the creation of this dummy started with: “primary activity last week”. The respondents were not asked directly whether they had a job or not. It is perfectly possible that e.g. some respondents were retired, but had a temporary job. This makes it thinkable that the surprising results of the regression are the results of discrepancy between the truth and the created dummy. Therefore, all that can be said is that there is no statistical evidence, but the parameters are correctly estimated, based on the dataset. These parameters will be kept.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
school_going	**HH**	1	0.4045	0.1155	12.2634	0.0005
school_going	**HO**	1	1.4820	0.0889	278.0721	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
school_going	**OH**	1	1.2595	0.0949	176.2056	<.0001
school_going	**OO**	1	1.2265	0.1026	142.8218	<.0001
school_going	**OW**	1	0.7008	0.1487	22.2102	<.0001
school_going	**WH**	1	0.2207	0.1096	4.0518	0.0441
school_going	**WO**	1	-0.2236	0.1742	1.6480	0.1992
school_going	**WW**	1	-1.0846	0.3261	11.0605	0.0009

Table 14: school going people parameters tour type model

This variable is comparable to the variable before because it also originates from the primary activity of last week. Besides the parameter for WO tours, they are all found to be significant. The reasoning of a discrepancy between the dummy and the truth is less likely because school going people can have a job in the evening or the weekend. Therefore, the parameters for work related trips can't be ignored. The parameter for tours going from work to work is negative and this makes sense, because the primary activity is going to school and they probably won't have a busy job. The probability of doing a not work related tour compared to HW tours is bigger for school going people than for working people. The variable for WH tours is not very big; indicating that if these respondents make a tour from work, they will likely also have a tour back starting at work.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
BEGINTIME_IN_MIN	**HH**	1	0.000136	0.000187	0.5319	0.4658
BEGINTIME_IN_MIN	**HO**	1	0.00266	0.000142	348.9051	<.0001
BEGINTIME_IN_MIN	**OH**	1	0.00458	0.000142	1042.8495	<.0001
BEGINTIME_IN_MIN	**OO**	1	0.00276	0.000178	241.0127	<.0001
BEGINTIME_IN_MIN	**OW**	1	0.00261	0.000207	158.3774	<.0001
BEGINTIME_IN_MIN	**WH**	1	0.00421	0.000156	725.6688	<.0001
BEGINTIME_IN_MIN	**WO**	1	0.00303	0.000188	258.9504	<.0001
BEGINTIME_IN_MIN	**WW**	1	-0.00064	0.000246	6.7758	0.0092

Table15: begin time parameters tour type model

Table15 shows the coefficients to calculate the difference in probability of a tour type with the reference (HW) caused by the begin time of a tour in minutes (to be calculated from 00:00). For every tour type

except HH tours there is a significant influence. This is purely statistical. A second model was constructed, using the OO tours as a reference. In this model, the HH tours have a significant difference because of “begin time in minutes”. Since this variable was calibrated with the same data as the variable “end time in minutes” (see above), it is normal that the effects are similar, because a tour with a larger begin time will also have a larger end time. The effects of begin time are smaller.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
DIST_M	**HH**	1	-0.0898	0.00201	1988.1611	<.0001
DIST_M	**HO**	1	0.00336	0.000469	51.1779	<.0001
DIST_M	**OH**	1	0.00344	0.000483	50.5263	<.0001
DIST_M	**OO**	1	0.00430	0.000474	82.2335	<.0001
DIST_M	**OW**	1	0.00304	0.000595	26.1813	<.0001
DIST_M	**WH**	1	0.00240	0.000531	20.5338	<.0001
DIST_M	**WO**	1	0.00206	0.000662	9.6827	0.0019
DIST_M	**WW**	1	-0.0502	0.00289	303.1875	<.0001

Table16: distance parameters tour type model

Dist_M is the distance of the longest tour segment in miles and hence, the distance of the main trip. However it is a mobility outcome, Davy Janssens (2005) stated that it is an important explaining variable for defining the trip purpose. There was referred to this research in chapter 7. Besides, there is a significant effect for every tour type (table 16). If the traveled distance gets longer, the probability of a HH or WW tour reduces compared to the reference HW tours. The other tour types have a higher probability when the distance grows compared to the reference. This is the same amount. Since many tours are work related, distance traveled can’t be seen independently from distance to work. This variable is discussed below.

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
GCDWORK	**HH**	1	0.00225	0.000340	43.5292	<.0001
GCDWORK	**HO**	1	0.00195	0.000331	34.6351	<.0001
GCDWORK	**OH**	1	0.00199	0.000337	34.6795	<.0001
GCDWORK	**OO**	1	0.00235	0.000336	48.8265	<.0001
GCDWORK	**OW**	1	0.00221	0.000368	35.9663	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	**TOURTYPE**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
GCDWORK	**WH**	1	0.000755	0.000401	3.5535	0.0594
GCDWORK	**WO**	1	0.00195	0.000370	27.7150	<.0001
GCDWORK	**WW**	1	0.00232	0.000370	39.2772	<.0001

Table17: great circle distance to work parameters tour type

GCDWORK is the great circle distance to work in miles. Its parameters are in table 17. WH tours are the only ones to have no significant effect. This parameter was kept because its insignificance is purely statistical: in the model using OO tours as a reference, there is a significant difference.

The parameters are small, but their corresponding values in the database have an average of 13 miles. It is thinkable that they will always be multiplied by this number. All signs are positive, indicating that the probabilities of any tour type grow relatively to the probability of a home work tour type when the great circle distance to work grows. They grow in approximately the same amount. This is an indication that people make fewer trips to their work (and corresponding from their work) if the distance to it grows. The time they win by this is more or less equally distributed over other tour types.

8.2 Distance model

This chapter describes the construction of a model that predicts the length in miles of a tour or trip (in the NHTS database every trip is treated as a tour, with an indicator if it actually was part of a tour) by using the variable “TOT_MILS”, which is the total distance of a tour in miles. It predicts this continuous variable and does not predict the chance of an outcome as in the tour type model.

The original model was built using the effect of weekday as a whole: a variable for every day of the week, keeping Wednesday as a reference. However, this seemed to be not significant, besides for the days Friday, Saturday and Sunday. This means that there is no statistical proof that trip miles on Monday, Tuesday and Thursday from trip miles on Wednesday and that distance is significantly different on Friday, Saturday and Sunday, compared to Wednesday. Therefore, a new model was built simplifying the weekday variable into a binary choice for week or weekend day. Now all variables are statistically significant. The first model is shown in table18, however it will be replaced by a better model. Table18 is shown for the reader’s convenience and understanding of how there was dealt with multicollinearity (see below).

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	17.13745	0.38252	44.80	<.0001	.	0
DISTTOWK	1	0.19924	0.00453	43.95	<.0001	0.93523	1.06926
GCDWORK	1	0.01818	0.00127	14.28	<.0001	0.95423	1.04796
HTPPOPDN	1	-0.00012419	0.00002514	-4.94	<.0001	0.61647	1.62215
HTEEMPDN	1	-0.00029283	0.00009310	-3.15	0.0017	0.54239	1.84369
HTHTNRNT	1	-0.02222	0.00565	-3.93	<.0001	0.69701	1.43470
urban	1	-1.61047	0.22366	-7.20	<.0001	0.81427	1.22810
CNTTDTR	1	-0.89238	0.03262	-27.35	<.0001	0.98775	1.01240
begntime_in_min_from_midnight	1	-0.14450	0.00117	-123.64	<.0001	0.07671	13.03694
endtime_in_min_from_midnight	1	0.14154	0.00116	121.76	<.0001	0.07674	13.03019
weekend	1	1.48538	0.18451	8.05	<.0001	0.99674	1.00327

Table18: distance model WITH MULTICOLINEARITY

Because some variables are likely to correlate with each other, there was calculated the variance inflation factor (VIF), along with its inverse the tolerance. The VIF is a measure that describes the phenomenon of multicollinearity; which occurs when a regressor is a linear combination of other regressors. The tollerance describes if this can safely be ignored (SAS). When looking at these values, it's obvious that the variables distance to work in miles (DISTTOWK), great circle distance to work in miles (GCDWORK), being urban (urban), count of travel day trips (CNTTDTR) and being in a weekend day (weekend) have small VIFs and great tolerance. Therefore, there are no problems of multicollinearity and they can be kept in the model. There are no problems of multicollinearity with the variables population per square mile (HTPPOPDN), workers per square mile living in the region (HTEEMPDN) and percentage renter occupied in the region (HTHTNRNT), however their VIFs are a little bit higher. Note that the parameters for these three variables are represented in the NHTS database (and hence, in the model) as the middlemost value of the real values (literally, not a median). This means that an area where the renter occupation is between 15% and 24%, the value in the model should be 20%! The other environmental variable (urban) has also a larger VIF and smaller tolerance than other variables, but this is not really problematically.

However, begin (begntime_in_min_from_midnight) and end time (endtime_in_min_from_midnight) correlate very highly with each other. There is assumed that their high multicollinearity is because these two variables both represent time. The tolerance is too low to keep both. they're both significant. This raises suspicion that it is rather the time when a tour takes place instead of the begin or end time. There was created a new variable: the mean of begin and end time (time_stamp). Since these both variables are calculated from midnight, this is the moment when a tour is in its half. The average deviation this

number gives from the start or begin time is only 13 minutes. This is a reason why the research prefers to create a new variable, although normally the averages of both variables should be subtracted of the individual values (center around the mean). There may be little information loss, but because begin and end time usually don't differ much, there is assumed that rather the time of day a trip is done than the begin and end times apart that influence the traveled distance. Another reason is because the difference between begin and end time influences the percentage of time spent on traveling and probably the traveled distance; longer trips lead to longer activities. This can be described with the concept travel time ratio (Vidacovic, 2000). Therefore, it is not really an interesting variable to put in the model. If these time stamps are found to be significant, this research assumes that it is safe to keep it.

The aim of this research is to find natural correlations between human and mobility variables. Count of travel day trips is rather a mobility outcome and correlates with the residence indicator urban/rural. Therefore, another model was built leaving out the variable CNTTDTR, in order to keep a simple model with only easily accessible input data.

Everything is ready to construct a new model, having no high multicollinearity and having only significant variables. This model is shown in table19.

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Tolerance**	**Variance Inflation**
Intercept	1	16.04607	0.35584	45.09	<.0001	.	0
DISTTOWK	1	0.23184	0.00462	50.20	<.0001	0.93953	1.06437
GCDWORK	1	0.01866	0.00130	14.36	<.0001	0.95425	1.04794
urban	1	-2.22294	0.22807	-9.75	<.0001	0.81630	1.22504
time_stamp	1	-0.00278	0.0003332 9	-8.35	<.0001	0.99772	1.00228
weekend	1	1.20482	0.18825	6.40	<.0001	0.99809	1.00192
Unique_Persons_HTPPOPDN	1	-0.0000879 8	0.0000256 6	-3.43	0.0006	0.61689	1.62103
Unique_Persons_HTEEMPDN	1	-0.0003653 0	0.0000950 5	-3.84	0.0001	0.54246	1.84346
Unique_Persons_HTHTNRNT	1	-0.02618	0.00577	-4.54	<.0001	0.69704	1.43464

Table19: distance model

It must be noted that the R-square value is 0.0104 (and the adjusted R-square value is the same), meaning that only 1% of the variability around its mean of the data is explained by the model. This is not

necessarily bad, because the model explains human behavior, which is hard to predict and will always have unexplainable variation (Frost, 2013). Besides, table 19 shows that all variables are (very) significant.

What remains, is a model with easy accessible and very significant variables. When a trip is made in a weekend, trips are on average 1,2 miles longer compared to a weekday. This may be because many people don't work in the weekends and hence, there is more time for longer trips. An urban tour is expected 2.4 miles shorter than tours from people who live in rural areas.

The intercept is very much bigger than any parameter. However, this does not mean that the model has less explanatory value since the values of most parameters can get quite high. The later on a day a tour takes place, the shorter it gets. The average of the variable time_stamp is 829 minutes from midnight. It is normal that later tours are shorter, because there is less time left. The parameter of population per square mile (HTPPOPDN) has a small absolute value and is negative, but its average value from the database is 3201. This average shortens a trip with 0,28 miles. The effect of renter occupancy (HTHTNRNT) in the area is bigger, but its values get smaller (an average of 26,2). This results on average in a decline of 0,68 miles due to renter occupancy. The variable workers per square mile (HTEEMPDN) has also a small effect, but a large average: 972, resulting in an average decline of 0.35 miles. The effects of these variables are consistent with the effect of the urban variable. It is thinkable that the population density and renter occupancy are higher in urban areas and that tours and trips get shorter.

Finally, two variables about the distance to work (DISTTOWK and GCDWORK) have an expectable influence; when the distance to work grows, the tours to work get longer and from other models (tour type model, section 8.1 and section 7.1), there is known that many tours and trips are work related. The average of distance to work is 12,8 miles and the average of great circle distance to work is 13,3 miles.

8.3 Activity duration model

This is a linear regression model that predicts the continuous variable activity duration in minutes. When a first run was made with the important variables found in section 7.4, it appeared to have little explaining value because the parameters of categorical variables (which could only be multiplied by one or zero) were too low and the intercept was too high; it just gave a basis for the average activity duration and the explaining variables didn't add much information; in every possible situation, the activity duration wouldn't have changed due to other variables. To solve this, the mean of every numerical variable was subtracted from its original value (center around the mean).

The variable CNTTDHH (number of household trips last week of the respondent's household) was deleted because it is seen rather a mobility outcome than an explaining variable. It is to be expected that it influences activity duration, but for the purpose of this research, it is something that is too difficult to investigate. It splits the research population in subgroups having other activity duration patterns, as was found in section 7.4, but the aim is to identify natural correlations between personal and household characteristics and mobility outcomes. This is a variable that may be hard to capture.

The variables begin time in minutes and end time in minutes (to be calculated from midnight) were in a first run found to be highly correlated. They were also very significant. The minutes traveled during the

tour also had an unacceptable collinearity with the other variables. There was already mentioned in this research that longer activities have a lower travel time ratio; the percentage of time spent on traveling is smaller (Vidacovic, 2000). Since begin and end times are calculated in minutes from midnight and minutes of travel was also calculated in minutes, it is thinkable that these three variables together cause the high multicollinearity. The begin and end times were replaced with a time stamp (the center of the time interval they contain). This is done in a similar way as in the traveled distance model (section 8.2).

When this was done and the model was constructed again, the problems with multicollinearity had disappeared. The model is in table 20.

Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Tolerance**	**Variance Inflation**
Intercept	1	207.40926	0.93082	222.82	<.0001	.	0
minutes_of_travel_minus_avg	1	0.79898	0.01639	48.73	<.0001	0.99503	1.00500
age_minus_avg	1	-0.50675	0.03085	-16.43	<.0001	0.98429	1.01596
flexibleworker	1	-40.13467	0.85040	-47.20	<.0001	0.97726	1.02327
weekday	1	56.28247	0.85047	66.18	<.0001	0.99955	1.00045
timestamp_in_min_minus_avg	1	-0.00204	0.0001689	-12.06	<.0001	0.99766	1.00234
parttimeworker	1	-38.91466	1.06069	-36.69	<.0001	0.96300	1.03842
multiplejobworker	1	-17.29127	4.08981	-4.23	<.0001	0.99664	1.00337
female	1	-11.80974	0.85762	-13.77	<.0001	0.95557	1.04649

table 20: activity duration model

The intercept is 207 minutes (approximately 3,45 hours). It was to be expected that longer travel times lead to longer activities. In section 7.4 this was already defended. There was found that non - daily activities have a lower ratio; a smaller percentage of the interval in which the activity took place was spent on traveling (Vidacovic, 2000). That means that the activity duration grows when the travel time grows.

Following this model, activities should on average get shorter with 0.5 minutes every year one gets older. The average value in the database is 47.5 years, which results in an expected decline of 23,75 minutes. The variable age was already mentioned to have an effect combined with other personal characteristics on mobility outcomes in the literature review (Jeppe Rich, 2010). The combined effect with gender was also mentioned. This model expects that women have about 12 minutes shorter activities than men.

Flexible workers (workers that can change their start or end time of work day) have shorter activities than non - flexible workers (about 40 minutes). It is possible that they perform more activities due to

their flexibility (e.g. work for half a day, do something else, work for two hours and do something else again) while fixed workers should spent more time without breaks at their work place. However there is no really indicator of number of trips a person did in a time period in the database, there is such an indicator for number of trips made by the whole household: CNTTDHH (number of household trips last week of the respondent's household). Flexible workers household have on average 12,8 weekly trips to 12,1 for non - flexible workers. This somehow defends this reasoning.

Activities on a weekday are on average 56 minutes longer than on a weekend day. This may be because of work activities, especially when they're combined with non-flexible workers. It means that on a weekday, other variables kept the same, an activity is more or less 4 hours, which can be half a work day (quartz.com, 2014). Respondents that leave the office, report the end of their activity. This reasoning can also be hold for part time workers who are less days at work. People with more than one job (multiplejobworker) have less activity duration reduction (17,3 minutes) than regular full time workers. Maybe they have more but smaller jobs and therefore more but shorter working activities or the reduction is due to less time availability. This remains unknown.

The time stamp is logic; the later it gets, the less time left for activities.

8.4 Mode choice model

This section describes the construction of a discrete choice model for mode choice on a trip. It will be a model that gives of the modal split between privately owned vehicles (cars), bicycle use (bic), walking (wal) or public transport (PT) use with privately owned vehicles as a reference. Some of the original categories were deleted to keep the model small and simple (see section 7.3). One of the variables that were found to influence this, is the household's life cycle (the NHTS uses this term, it is something like the household's composition). The NHTS uses this code:

- 01 = one adult, no children
- 02 = 2 or more adults, no children
- 03 = one adult, youngest child 0-5
- 04 = 2 or more adults, youngest child 0-5
- 05 = one adult, youngest child 6-15
- 06 = 2 or more adults, youngest child 6-15
- 07 = one adult, youngest child 16-21
- 08 = 2 or more adults, youngest child 16-21
- 09 = one adult, retired, no children
- 10 = 2 or more adults, retired, no children

This was simplified into these categories to keep the model small:

- Bachelor (codes 01 and 09)
- Two adults (with grownups) (code 02, 07, 08 and 10)
- Regular family (codes 04 and 06)
- One adult with a minor (codes 03 and 05)

In the model, the reference for this variable will be a “regular family”.

Another categorical variable, is education level. The code is:

- 01 = less than high school graduate
- 02 = high school graduate
- 03 = some college or Associate's degree
- 04 = bachelor's degree
- 05 = graduate or Professional Degree

The reference will be college or associate’s degree.

Other variables that are expected to have a significant effect are: whether there are any stops, begin time, housing units per square mile, vehicle count, total distance traveled and time traveled.

The model results are in tables 21 to 32. Table 21 shows the intercept estimates.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	**PT**	1	-2.6205	0.0717	1334.3878	<.0001
Intercept	**bic**	1	-4.0856	0.1044	1531.7918	<.0001
Intercept	**wal**	1	-2.8510	0.0329	7527.7491	<.0001

Table 21: intercept mode choice model

As can be seen, every mode has a negative intercept; they start with a smaller probability towards car use. This is not surprising, 92% of the reported trips are done with a car. The estimates are very significant.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Bachelor	**PT**	1	-0.3398	0.0517	43.2481	<.0001
Bachelor	**bic**	1	-0.1924	0.0780	6.0929	0.0136
Bachelor	**wal**	1	0.0732	0.0250	8.5909	0.0034

Table22: bachelor parameters mode choice model

Table 22 shows the parameters that modify the probability of bachelors to use other modes. Bachelors have a little bigger probability to prefer walking above car use than regular families. The effect is however very small. The probability that they reject the choices of public transport and bicycle use is much bigger. This effect is normal, because most bachelors do have a car (see figure 19).

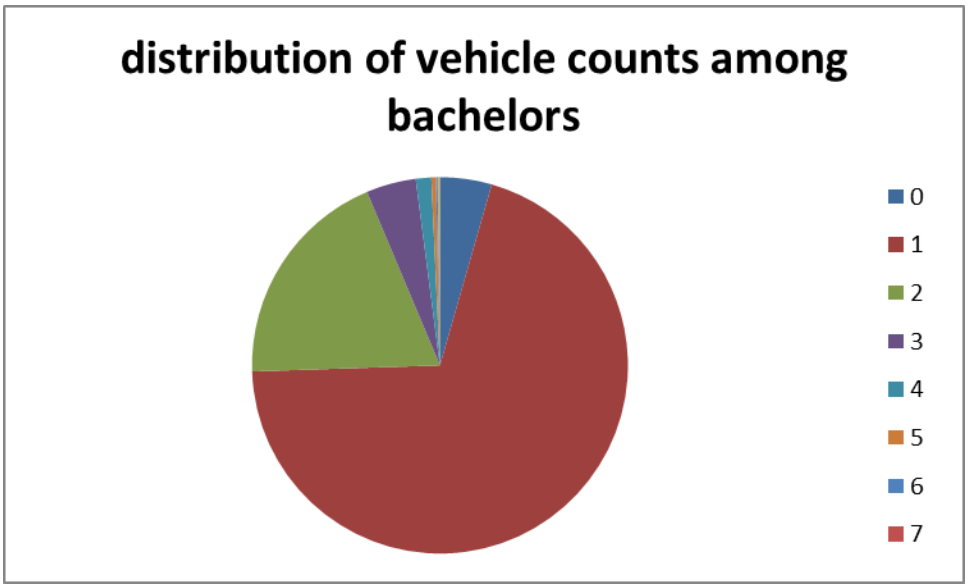

Figure 19: distribution of vehicle counts among bachelors

Since these people live alone, there's a smaller chance that it is not available compared to households with more car users. Therefore, they feel they are better off using the car (Cees Wildervanck, 1996).

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Two_adults with grownups	**PT**	1	0.1458	0.0347	17.7054	<.0001
Two_adults with grownups	**bic**	1	0.000228	0.0458	0.0000	0.9960
Two_adults with grownups	**wal**	1	0.1080	0.0145	55.4038	<.0001

Table23: two adults households parameters mode choice model

It stands out that there is no proven difference of bicycle use towards car use by two adults households. However, there is a significant difference when the reference was set to bachelor.

Two adults have a bigger probability to choose public transport and walking compared to car use than regular families. It may be that these households are younger and live in urban areas, where these modes are easier to access. It is a fact that a higher percentage of younger people live in urban areas (Pateman, 2011). This reasoning also holds for the positive sign of bicycle use.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
One_adult_minor	**PT**	1	-0.2286	0.0986	5.3745	0.0204
One_adult_minor	**bic**	1	-0.7655	0.2050	13.9413	0.0002
One_adult_minor	**wal**	1	0.0561	0.0504	1.2384	0.2658

table 24: one adult and minors household parameters mode choice model

Table 24 shows the estimates for households with one adult and a minor. There is no statistical evidence of a difference with regular families in walking as mode choice, but when the reference was changed to bachelors, there was. It may be that walking is perceived safe for a minor or that these people are younger and live in the city where more destinations are on walk distance (Pateman, 2011). This conflicts with the negative signs of the parameters for public transport and bicycle. These adults have young children and hence, have higher car availability if they own one (no one else can be gone using it). Besides, the car is perceived as a safe social mode (Peeters, 2000).

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
less_than_highschool	**PT**	1	0.4309	0.0656	43.0980	<.0001
less_than_highschool	**bic**	1	0.3284	0.1069	9.4361	0.0021
less_than_highschool	**wal**	1	0.1323	0.0394	11.2746	0.0008

Table25: less than highschool education parameters mode choice model

Table 25 shows the parameters for mode choice of people with a lower grade than high school (in the American sense of the word, not European since the NHTS comes from America) compared to the reference, “people with a college or associates degree”. All estimates are significant. It is noticeable that all the signs are positive, meaning that the probability of public transport, bicycle use or walking increases if one is in this demographic category. It is because these modes are cheaper and lower educated people have less income (Rawlinson, 2011). This is illustrated in figure 20. Influence of income was also confirmed by Ibrahim Sheikh A. K. (2007).

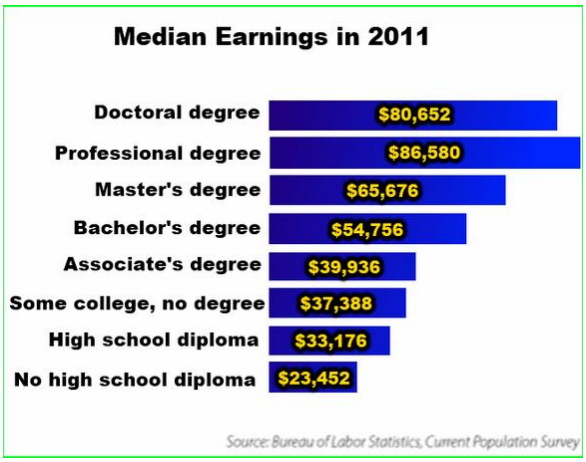

Figure 20: median earnings in 2011 (source: Bureau of Labor Statistics)

Analysis of Maximum Likelihood Estimates						
Parameter	mode	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
high_school_graduate	PT	1	-0.1722	0.0464	13.7591	0.0002
high_school_graduate	bic	1	-0.1525	0.0703	4.7085	0.0300
high_school_graduate	wal	1	-0.2045	0.0221	85.5617	<.0001

table 26: high school graduate parameters mode choice model

Table 26 shows the parameters for people having a high school graduate. Since it is the second lowest category, it is still not very high. However, these people earn averagely 10000$ a year more (Rawlinson, 2011), which enables them to afford other modes. This explains why they have negative signs and lower educated people have positive signs.

Analysis of Maximum Likelihood Estimates						
Parameter	mode	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
bachelors_degree	PT	1	-0.3295	0.0459	51.4689	<.0001
bachelors_degree	bic	1	0.2030	0.0608	11.1323	0.0008
bachelors_degree	wal	1	0.2321	0.0188	153.1240	<.0001
graduate_or_professi	PT	1	-0.1217	0.0452	7.2442	0.0071
graduate_or_professi	bic	1	0.4810	0.0596	65.2224	<.0001
graduate_or_professi	wal	1	0.4881	0.0187	679.8925	<.0001

Table27: bachelor degrees and professional degrees parameters mode choice model

People with a bachelor degree and people with a professional degree are discussed together because they have a similar mode choice pattern.

It is strange that bicycle use and walking have a bigger probability than car use for bachelor degrees than for people with a college degree (Table 27). It may be that, because of their higher income (figure 20), have more freedom to choose these modes and that they are considered recreational. This would also explain that they don't prefer public transport. This effect is larger when people are graduated, maybe because of the income differences.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
intermediate_stops	**PT**	1	-0.5396	0.0453	141.6938	<.0001
intermediate_stops	**bic**	1	-0.7609	0.0759	100.4214	<.0001
intermediate_stops	**wal**	1	-1.0295	0.0249	1705.1713	<.0001

Table28: intermediate stops parameters mode choice model

Table 28 shows that car is preferred over other modes if there are intermediate stops (note that it is just an indicator if there are any, not HOW many). This is probably because a car offers flexibility. The big absolute value for walking may be because tours with stops may be too long to walk.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
TOT_MILS	**PT**	1	-0.0152	0.000653	538.7440	<.0001
TOT_MILS	**bic**	1	-0.0833	0.00364	523.7560	<.0001
TOT_MILS	**wal**	1	0.000728	0.000094	60.4391	<.0001
TOT_CMIN	**PT**	1	0.0178	0.000416	1837.5643	<.0001
TOT_CMIN	**bic**	1	0.0171	0.000475	1292.5113	<.0001
TOT_CMIN	**wal**	1	0.00164	0.000233	49.7495	<.0001

Table29: total miles and total minutes parameters mode choice model

Table 29 shows the estimates for the related variables of the total minutes of travel and the total miles of a trip. The results are significant but unexpected. At first, they have opposite signs for public transport and bicycling. Secondly, they both show that how longer a trip gets (in miles or minutes) the probability of walking increases. The parameters are all low, but may be multiplied with quite high values (average of TOT_MILS is 14,5 and TOT_CMIN is 27). It is thinkable that because of this, the probability of walking stays low because its parameter is at least for both parameters ten times lower than the parameters for

public transport and bicycling. The opposite signs for TOT_MILS and TOT_CMIN may be because of leisure trips (much time) and professional or necessary trips (many miles).

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
HHVEHCOUNT	**PT**	1	-0.8536	0.0211	1643.6067	<.0001
HHVEHCOUNT	**bic**	1	-0.3535	0.0253	194.7912	<.0001
HHVEHCOUNT	**wal**	1	-0.2193	0.00749	857.1578	<.0001

Table 30: HH vehicle count parameters mode choice model

Table 30 shows how the probabilities of public transport, bicycling and walking changes when the vehicle count of household changes. All estimates are very significant and have a negative sign, which is normal because when there are more vehicles available this is more often in people's modes choice set. The parameter for public transport has more influence than the ones for bicycling and walking. This may be because the latter two can be used for leisure.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
HTRESD	**PT**	1	0.000145	2.318E-6	3925.6017	<.0001
HTRESD	**bic**	1	0.000056	5.591E-6	99.6440	<.0001
HTRESD	**wal**	1	0.000105	1.782E-6	3455.5606	<.0001

Table31: house density parameters mode choice model

Housing units per square mile has a positive effect on the probability of public transport, bicycling and walking towards car use (table 31). It was argued in this model before that people living in cities may use these modes more often. This corresponds with the positive signs of housing units per square mile, which is likely to be greater in cities than in rural areas. Note that the average in the database is 1379 houses per square mile, so the relative small parameters can result in large probability changes.

Analysis of Maximum Likelihood Estimates						
Parameter	**mode**	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
begintime_in_min	**PT**	1	-0.00066	0.000057	135.4669	<.0001
begintime_in_min	**bic**	1	-0.00001	0.000078	0.0318	0.8584
begintime_in_min	**wal**	1	0.000481	0.000024	401.1010	<.0001

Table32: begintime parameters mode choice model

Table 32 shows the influence of begin time in minutes from midnight. For bicycle, there is no significant difference compared to private vehicles. However, when the model was constructed again with using walking as reference mode, there was a significant difference. Note that the average values of begin time is 728 minutes, so that it can have a relevant influence. The signs for bicycle and public transport are as expected: if it gets later, weather circumstances may affect bicycle use and public transport is not available when it is late. The positive sign for walking trips may be because of short, walkable trips when it is later (leisure).

9. Conclusion, discussion and further work

This thesis started to describe difficulties in data capturing to discuss mobility behavior. It stated that well substantiated correlations can be a help for future decision makers. Next, indications were sought in literature on what stable correlations might exist. These were tested by using decision trees. After that, they were transformed to mathematical models that describe these correlations quantitatively. This chapter will summarize personal characteristics that often come back and hence, are stable natural indicators for mobility behavior. These can be influenced to change this behavior. This chapter brings together all the models without proving a statistical correlation. It can be a source of further work.

A first factor appearing in many models is the factor time of day. It has influence on the tour type, mode choice, duration and distance. This relationship seems to be obvious. It can be an opportunity for further research to build a day covering model predicting these four outcomes (and perhaps of other ones) simultaneously. This research can conclude that there is a different behavior during the week than during the weekend.

Another very important factor that keeps coming back, is work in different aspects. Many tours include a work destination. The tour type model predicts that a tour going from work to home has a smaller probability to occur. This may be because people do more stops when they come back from work (again a factor time). Another aspect of work is the work status. One's work situation (active or non-active) and one's type of work, which should be partially defined by educational level (this appears in the mode choice model). With these two facts combined, there can be stated that further research on work trips can provide insight in mode choice, because stops are likely to be when people return from work. This may also be a reason not to use public transport, following the mode choice model.

Work situation also returns in the activity duration model, where the difference between fulltime, part time and multiple job workers was investigated.

In the two paragraphs above, it becomes clear that work situation in different aspects is one of the biggest factors that influences mobility. Therefore, further research to build a model predicting work trips as a whole (mode choice, distance, stops …) can be useful. Distance to work is also something that comes back in the distance model and the tour type model.

A third factor that reappears is the environment one lives in (urban versus rural, population density, working density …). This influences mode choice and traveled distance. However, it does not influence activity duration. The concept of travel time ratio was more than once discussed in this research

(Vidacovic, 2000). It states that longer trips lead to longer activities: the percentage of time traveled of the whole activity duration (including travel time) gets smaller. This raises suspicion that people who live in rural areas also have longer activities, although this did not come forward in this research.

An overall conclusion of this research is that the most important personal and household characteristics are not a characteristic of the individual (age, gender …). They don't have a very big influence and don't come back repeatedly. The characteristics that influence mobility behavior are rather aspects of the live of respondents: work, living place and time of day. There must be noted that here are only four mobility outcomes investigated. Other ones and combinations that are proposed here above may have a bigger influence. This is not done in this research, but can be done in the future.

A last thing to consider is if these models and trends are transferable to other regions. The National Household Travel Survey was carried out in the United States of America. It is not self-evident that people in other parts of the world have the same behavior. Here is made a comparison between the USA and Europe, to see if the results hold in the latter region.

The gross domestic product (GDP) per capita in the Euro Area (composed of 18 countries: Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Luxembourg, Malta, Netherlands, Portugal, Slovak Republic, Slovenia and Spain) equals 11681,147$. (International monetary fund, 2015). The GDP per capita of the USA is higher, $54629,5 (Worldbank.org, 2014). If the average GDP per capita of the top twenty richest countries of Europe is calculated, it is 62367,55$ (25 Richest Countries in Europe, 2012). It is the average of the GDPs of Monaco, Liechtenstein, Luxembourg, Norway, Switzerland, San Marino, Denmark, Sweden, Austria, The Netherlands, Ireland, Iceland, Belgium, Germany, United Kingdom, Andorra, Finland, France, Italy and Spain. This GDP per capita is higher, but since it is average and there is nothing known about the distribution in the USA, there is decided that the richer regions of Europe and the USA can be compared with each other for income. Work situation is one of the most important factors defining mobility, as was stated here above. Income has at least to do with work situation (however not every aspect). Therefore, this research concludes that there is no reason to reject that the results are transferable to the richer lands of Europe. However, it may be interesting for further research to investigate the effects of culture on mobility behavior.

Further, the NHTS holds some typical attributes that characterizes the ethnical background of a respondent (like race, born in the USA, Hispanic roots…). They were never found to be significant. The research concludes that it is safe to transfer the data to European countries.

Bibliography

D. Janssens, K. D. (2014). *OVG Vlaanderen 4.5.*

Davy Janssens, G. W. (2005). *Integrating Bayesian networks and decision trees in a sequential rule-based transportation model.*

Ibrahim Sheikh A. K., R. U. (2007). *Mode Choice Model for Vulnerable Motorcyclists in.*

Jeppe Rich, C. G. (2010). *ACTIVITY-BASED DEMAND MODELLING ON A LARGE SCALE: EXPERIENCE FROM THE NEW DANISH NATIONAL MODEL.*

Linda Nijland, T. A. (2014). *5. Multi-day activity scheduling reactions to planned activities and future events in a dynamic model of activity-travel behavior.*

Maine. (2015, april 19). Opgehaald van Labor Market Area Definitions: http://www.maine.gov/labor/cwri/LMADefinitions.html

Renni Anggraine, T. A. (2010). 4. Modeling household activity participation decisions in a rule-based system of travel demand. *Journal of the Eastern Asian Society for Transportation Studies*.

Shiftan, Y. (2008). *7. The use of activity-based modeling to analyze the effect of land-use policies on travel behavior.*

Theo A. Arentze, H. J. (2002). *A learning-based transportation oriented simulation system.*

Zhihu Zhanga, H. G. (2013). *A Traffic Mode Choice Model for the Bus User Groups Based on SP and RP Data.*

(2014). Retrieved from quartz.com: http://qz.com/437121/the-9-to-5-office-workday-is-dying-in-america/

Jeppe Rich, C. G. (2010). *ACTIVITY-BASED DEMAND MODELLING ON A LARGE SCALE: EXPERIENCE FROM THE NEW DANISH NATIONAL MODEL.*

Vidacovic, M. D. (2000). *Travel Time Ratio: the key factor to spatial reach.*

SAS. (n.d.). *Collinearity Diagnostics*. Retrieved 08 11, 2015, from support.sas.com: http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect038.htm

Cees Wildervanck, G. T. (1996). *Autogebruik te sturen?* Adviesdienst Verkeer en Vervoer.

Ibrahim Sheikh A. K., R. U. (2007). *Mode Choice Model for Vulnerable Motorcyclists in.*

Pateman, T. (2011). *Rural and urban areas: comparing lives using rural/urban classifications.* Office for National Statistics.

Peeters, K. (2000). *Het Voor(r)uitperspectief.*

Rawlinson, K. (2011). *Is Education Important?* Retrieved from http://eclecticsite.com/: http://eclecticsite.com/education&income.html

OECD. (2014). *Average annual hours actually worked per worker*. Retrieved August 17, 2015, from OECD: https://stats.oecd.org/Index.aspx?DataSetCode=ANHRS

Schawbell, D. (2011, December 21). *The Beginning of the End of the 9-to-5 Workday?* Retrieved August 15, 2015, from Business.Time.com: http://business.time.com/2011/12/21/the-beginning-of-the-end-of-the-9-to-5-workday/

McGuckin, N. (2004). *Trips, Chains, and Tours—Using an Operational Definition.* Retrieved June 30, 2015, from online pubs: http://onlinepubs.trb.org/onlinepubs/archive/conferences/nhts/mcguckin.pdf

(2000). *OVG Antwerpen.*

(2001). *OVG Gent.*

World Bank, GDP Per capita. (n.d.). Retrieved 07 31, 2015, from data of world bank: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD

How is rural defined? (n.d.). Retrieved 07 25, 2015, from HRSA: http://www.hrsa.gov/healthit/toolbox/RuralHealthITtoolbox/Introduction/defined.html

(2014). Retrieved from Worldbank.org: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD/countries/1W?display=default

25 Richest Countries in Europe. (2012). Retrieved 2015, from http://www.listnbest.com/25-richest-countries-europe-wealthiest-european-countries/

International monetary fund. (2015). Retrieved from http://www.imf.org/external/pubs/ft/weo/2015/01/weodata/weoselagr.aspx

Vidacovic, M. D. (2000). *Travel Time Ratio: the key factor to spatial reach.*

Appendix 1: attributes National Household Travel Survey

Attribute	Description	Attribute	Description
HOUSEID	HH eight-digit ID number	RAIL	MSA heavy rail status for HH
VARSTRAT	Linearization Variance Stratum for Std Err Calculation	RESP_CNT	Count of responding persons per HH
WTHHFIN	Final HH weight	SCRESP	Person ID number of screener respondent
DRVRCNT	Number of drivers in HH	TRAVDAY	Travel day
CDIVMSAR	Grouping of HH by combination of Census Division, MSA status, and presence of a subway system (if area > 1 million)	URBAN	Home address in urbanized area
CENSUS_D	Census division classification for home address	URBANSIZE	Size of urban area in which home address is located
CENSUS_R	Census region classification for home address	URBRUR	Household in urban/rural area
HH_HISP	Hispanic status of HH respondent	WRKCOUNT	Number of workers in HH
HH_RACE	Race of HH respondent	TDAYDATE	Travel day date
HHFAMINC	Derived total HH income	FLAG100	Did HH have 100% of members complete interview?
HHRELATD	At least some HH members are related	LIF_CYC	Life Cycle classification for the HH
HHRESP	Person ID number of household respondent	CNTTDHH	Category of number of household trips
HHSIZE	Count of HH members	HBHUR	Urban / Rural indicator - Block group
HHSTATE	State HH location	HTRESDN	Housing units per sq mile - Tract level
HHSTFIPS	State FIPS for HH address	HTHTNRNT	Percent renter-occupied - Tract level
HHVEHCNT	Count of HH vehicles	HTPPOPDN	Population per sq mile - Tract level
HOMEOWN	Housing unit owned or rented	HTEEMPDN	Workers per square mile living in Tract
HOMETYPE	Type of housing unit	HBRESDN	Housing units per sq mile - Block group

MSACAT	MSA category for the HH home address	HBHTNRNT	Percent renter-occupied - Block group
MSASIZE	MSA population size for the HH home address	HBPPOPDN	Population per sq mile - Block group
NUMADLT	Count of adult HHMs at least 18 years old	HH_CBSA	CBSA FIPS code for HH address
		HHC_MSA	CMSA FIPS code for HH address

Table A: Attributes HH file

Attribute	Attribute Description	Attribute	Attribute Description	Attribute	Attribute Description
HOUSEID	HH eight-digit ID number	TRVL_MIN	Derived trip time - minutes	USEPUBTR	Use public transit on travel day
PERSONID	Person ID number	TRVLCMIN	Calculated travel time	VEHID	HH vehicle number used for trip
FRSTHM	Did Person Start Travel Day at Home?	TRWAITTM	Derived length of wait for public transit - minutes	WHODROVE	Person ID of driver on trip
OUTOFTWN	R Was out of town the entire travel day	INTSTATE	A part of this trip was on interstate	WHYFROM	Trip purpose for previous trip
ONTD_P1	Person number 1 was on travel day trip	GASPRICE	Price of gasoline (cents) on respondent's travel day	WHYTO	Travel day purpose of trip
ONTD_P2	Person number 2 was on travel day trip	VEHTYPE	Vehicle type	WHYTRP1S	Trip purpose summary
ONTD_Pn (3 to 15)	Person number n was on travel day trip	NONHHCNT	Derived number of non-HHMs on trip	WRKCOUNT	Number of workers in HH
TDCASEID	Trip number	NUMONTRP	Count of total people on trip	DWELTIME	Calculated Time (minutes) at Destination
DRIVER	Driver status of S	PAYTOLL	Toll paid on this interstate	WHYTRP90	1990 Trip Purpose
R_SEX	Respondent gender	PRMACT	Primary activity last week	TDTRPNUM	Travel Day Trip number

WORKER	Subject worker status	PROXY	Trip info from respondent or proxy	TDWKND	Travel day trip was on weekend
TRIPPURP	General Trip Purpose (Home-Based Purpose types)	PSGR_FLG	S was passenger on trip that only used POV (privately owned vehicle)	TREGRn (n = 1 to 5)	nth mode used from public transit to destination
AWAYHOME	Travel day reason subject was away from home	R_AGE	Respondent age	TRPMILES	Calculated Trip distance converted into miles
DROP_PRK	Parked or dropped off at public transit	USEINTST	Interstate used for any trips	WTTRDFIN	Final trip weight
DRVR_FLG	Subject was driver on this trip	STRTTIME	Start time	VMT_MILE	Calculated Trip distance (miles) for Driver Trips
EDUC	Highest grade completed	TRACCn (n = 1 to 5)	nth mode used to get to public transit	PUBTRANS	Respondent Used Public Transportation on trip
ENDTIME	Trip_end_time	TRACCTM	Derived time to get to public transit - minutes	TRPHHACC	Number of HHM with R on trip
HH_ONTD	Derived number of HH members on trip	TREGRTM	How long to destination from transit - converted to minutes	TRPHHVEH	Number of HHM with R on trip
HHMEMDRV	HH member drove on trip	TRPACCMP	Number of people with R on trip	TRPTRANS	Transportation mode used on trip (as reported by respondent)

Table B: person trip characteristics

All attributes:
HOUSEID,PERSONID,VARSTRAT,WTPERFIN,SFWGT,HH_HISP,HH_RACE,DRVRCNT,HHFAMINC,HHSIZE,HHVEHCNT,NUMADLT,WRKCOUNT,FLAG100,LIF_CYC,CNTTDTR,BORNINUS,CARRODE,CDIVMSAR,CENSUS_D,CENSUS_R,CONDNIGH,CONDPUB,CONDRIDE,CONDRIVE,CONDSPEC,CONDTAX,CONDTRAV,DELIVER,DIARY,DISTTOSC,DRIVER,DTACDT,DTCONJ,DTCOST,DTRAGE,DTRAN,DTWALK,EDUC,EVERDROV,FLEXTIME,FMSCS

IZE,FRSTHM,FXDWKPL,GCDWORK,GRADE,GT1JBLWK,HHRESP,HHSTATE,HHSTFIPS,ISSUE,OCCAT,LSTTRDAY,MCUSED,MEDCOND,MEDCOND6,MOROFTEN,MSACAT,MSASIZE,NBIKETRP,NWALKTRP,OUTCNTRY,OUTOFTWN,PAYPROF,PRMACT,PROXY,PTUSED,PURCHASE,R_AGE,R_RELAT,R_SEX,RAIL,SAMEPLC,SCHCARE,SCHCRIM,SCHDIST,SCHSPD,SCHTRAF,SCHTRN1,SCHTRN2,SCHTYP,SCHWTHR,SELF_EMP,TIMETOSC,TIMETOWK,TOSCSIZE,TRAVDAY,URBAN,URBANSIZE,URBRUR,USEINTST,USEPUBTR,WEBUSE,WKFMHMXX,WKFTPT,WKRMHM,WKSTFIPS,WORKER,WRKTIME,WRKTRANS,YEARMILE,YRMLCAP,YRTOUS,DISTTOWK,TDAYDATE,HOMEOWN,HOMETYPE,HBHUR,HTRESDN,HTHTNRNT,HTPPOPDN,HTEEMPDN,HBRESDN,HBHTNRNT,HBPPOPDN,HH_CBSA,HHC_MSA.

attribute	attribute description	attribute	attribute description	attribute	attribute description
WTPERFIN	Final person weight	FMSCSIZE	Number of people on from school trip	SCHSPD	Walk/Bike issue: speed of traffic along route
SFWGT	Weight for child 5-15, Safe Routes to School section	FXDWKPL	No fixed workplace	SCHTRAF	Walk/Bike issue: speed of traffic along route
CNTTDTR	Count of travel day trips for this respondent	GCDWORK	Great circle distance (miles) between home and work	SCHTRN1	Mode to school
BORNINUS	Resp born in US?	GRADE	Grade allowed to walk/bike to/from school without adult	SCHTRN2	Mode from school
CARRODE	Number of people in vehicle last week	GT1JBLWK	Have more than one job	SCHTYP	School type
CONDNIGH	Medical condition results in limiting driving to daytime	ISSUE	Most important transportation issue	SCHWTHR	Walk/Bike issue: poor weather or climate in area
CONDPUB	Medical condition results in using bus/subway less frequently	OCCAT	Job category	SELF_EMP	Self-employed
CONDRIDE	Medical condition results in asking others for rides	LSTTRDAY	Approximate number of days since last trip	TIMETOSC	Minutes to get to school
CONDRIVE	Medical condition results in giving up driving	MCUSED	Times used motorcycle/moped on road in the past month	TIMETOWK	Minutes to go from home to work last week
CONDSPEC	Medical condition results in using special transit services	MEDCOND	Have medical condition making it hard to travel	TOSCSIZE	Number of people on to school trip
CONDTAX	Medical condition	MEDCOND6	Length of time with	WEBUSE	Frequency of

	results in using a reduced fare taxi		medical condition		internet use in past month
CONDTRAV	Medical condition results in reduced day-to-day travel	MOROFTEN	Would like to get out more often	WKFMHMXX	Frequency of working from home in past month
DELIVER	Number of these internet purchases delivered to home	NBIKETRP	Number of bike trips in past week	WKFTPT	Work full or part-time
DIARY	Indicates if travel diary was completed	NWALKTRP	Number of walk trips in past week	WKRMHM	Has option to work at home
DISTTOSC	Distance home to school (miles)	OUTCNTRY	S out of country entire travel day	WKSTFIPS	State FIPS code for work address
DTACDT	Respondent's view on Safety concerns	PAYPROF	Worked for pay or profit last week	WRKTIME	Usual arrival time at work
DTCONJ	Respondent's view on Highway congestion	PTUSED	How often S used public transit in past month	WRKTRANS	Transportation mode to work last week
DTCOST	Respondent's view on Price of travel (fees, tolls and gas)	PURCHASE	Number of times purchased via internet in past month	YEARMILE	Miles respondent drove last 12 months
DTRAGE	Respondent's view on Aggressive/distracted drivers	R_RELAT	Respondent relationship to HH respondent	YRMLCAP	Indicates YEARMILE was capped
DTRAN	Respondent's view on Access or availability of public transit	SAMEPLC	Stayed at same place all day	YRTOUS	Year entered U.S.
DTWALK	Respondent's view on Lack of walkways or sidewalks	SCHCARE	Attends before or after school care	DISTTOWK	One-way distance to workplace (miles)
EVERDROV	Has been a driver in the past	SCHCRIM	Walk/Bike issue: violence/crime along route		
FLEXTIME	Respondent can set or change start time of work day	SCHDIST	Walk/Bike issue: distance between home & school		

Table C: person file attributes

Variable	Type	Len	Label
BEGNTIME	Char	4	Tour begin time (HHMM)
DIST_M	Num	8	Distance longest segment (miles)
ENDTTIME	Char	4	Tour end time (HHMM)

HOUSEID	Char	8	HH eight-digit ID number
MODE_D	Char	2	Mode of longest distance segment
MODE_T	Char	2	Mode of longest time segment
PERSONID	Char	2	Person ID
PMT_OTHR	Num	8	Tour level PMT for other modes
PMT_POV	Num	8	Tour level PMT for POV
PMT_TRAN	Num	8	Tour level PMT for transit
PMT_WALK	Num	8	Tour level PMT for walk
STOPS	Num	8	Number of stops for the tour
TIME_M	Num	8	Time of longest segment (minutes)
TOT_CMIN	Num	8	Tour level calculated minutes of travel
TOT_DWEL	Num	8	Tour level dwell times for intermediate stops (minutes)
TOT_DWEL2	Num	8	Tour level dwell times for all stops (minutes)
TOT_MILS	Num	8	Tour level total miles of travel
TOUR	Num	8	Sequential tour number for person (1-N)
TOURTYPE	Char	2	Type of tour
TOUR_FLG	Char	1	1=Yes Part of Tour, 0=Not Tour
VMT	Num	8	Tour level VMT
WTTRDFIN	Num	8	Final trip weight

Table D: attributes tour file (copied from NHTS manuals)

Variable	Type	Len	Label
HOUSEID	Char	8	HH eight-digit ID number
PERSONID	Char	2	Person ID number
STOPS	Num	8	Number of stops for the tour
TDTRPNUM	Char	12	Travel Day Trip number
TOUR	Num	8	Sequential tour

			number for person (1-N)
TOURTYPE	Char	2	Type of tour
TOUR_FLG	Char	1	1=Yes Part of Tour, 0=Not Tour
TOUR_SEG	Num	8	Sequential location of trip within tour (1-N)
TRPCNT	Num	8	Number of trips that make up the tour
WTTRDFIN	Num	8	Final trip weight

Table E: attributes chaintrip file (copied from NHTS manuals)

Appendix 2: decision tree distance class

Correctly Classified Instances	48880	51.4288 %
Incorrectly Classified Instances	46164	48.5712 %
Kappa statistic	0.3779	
Mean absolute error	0.0985	
Root mean squared error	0.2221	
Relative absolute error	79.6446 %	
Root relative squared error	89.3105 %	
Total Number of Instances	95044	

```
DISTTOWK <= 10.5545
|   DISTTOWK <= 3.333
|   |   TOUR_FLG <= 0: A1 (75178.0/24440.0)
|   |   TOUR_FLG > 0
|   |   |   USEINTST <= 1
|   |   |   |   STOPS <= 1: A1 (2694.0/2170.0)
|   |   |   |   STOPS > 1: B (1481.0/1119.0)
|   |   |   USEINTST > 1
|   |   |   |   STOPS <= 1
|   |   |   |   |   DISTTOWK <= 2.5553: A1 (4941.0/2818.0)
|   |   |   |   |   DISTTOWK > 2.5553: A2 (2047.0/1279.0)
|   |   |   |   STOPS > 1
|   |   |   |   |   DISTTOWK <= 1.9998: A1 (1764.0/1314.0)
|   |   |   |   |   DISTTOWK > 1.9998: A2 (2101.0/1530.0)
|   DISTTOWK > 3.333
|   |   DISTTOWK <= 6.1105
|   |   |   TOUR_FLG <= 0
|   |   |   |   STOPS <= 0
|   |   |   |   |   BEGNTIME <= 854: A2 (10322.0/2375.0)
|   |   |   |   |   BEGNTIME > 854
|   |   |   |   |   |   CNTTDTR <= 2: A2 (5334.0/1559.0)
|   |   |   |   |   |   CNTTDTR > 2
|   |   |   |   |   |   |   Unique_Persons.HBPPOPDN <= 300: A2 (7418.0/4055.0)
|   |   |   |   |   |   |   Unique_Persons.HBPPOPDN > 300
|   |   |   |   |   |   |   |   CARRODE <= -1: A1 (1644.0/887.0)
|   |   |   |   |   |   |   |   CARRODE > -1
|   |   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY <= 1: A1 (2907.0/1856.0)
|   |   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY > 1
|   |   |   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY <= 6
```

```
|  |  |  |  |  |  |  |  |  |  ENDTTIME <= 1819: A2 (11521.0/5496.0)
|  |  |  |  |  |  |  |  |  |  ENDTTIME > 1819
|  |  |  |  |  |  |  |  |  |  |  CNTTDTR <= 5: A2 (1994.0/1139.0)
|  |  |  |  |  |  |  |  |  |  |  CNTTDTR > 5: A1 (2455.0/1503.0)
|  |  |  |  |  |  |  |  |  Unique_Persons.TRAVDAY > 6: A1 (3093.0/1957.0)
|  |  |  |  STOPS > 0: A1 (5840.0/3026.0)
|  |  |  TOUR_FLG > 0
|  |  |  |  USEINTST <= 1
|  |  |  |  |  STOPS <= 1
|  |  |  |  |  |  CNTTDTR <= 6: B (1136.0/892.0)
|  |  |  |  |  |  CNTTDTR > 6: A3 (1207.0/912.0)
|  |  |  |  |  STOPS > 1: B (1208.0/875.0)
|  |  |  |  USEINTST > 1
|  |  |  |  |  STOPS <= 1: A2 (5143.0/3107.0)
|  |  |  |  |  STOPS > 1
|  |  |  |  |  |  Unique_Persons.HTRESDN <= 750: B (1585.0/1050.0)
|  |  |  |  |  |  Unique_Persons.HTRESDN > 750: A3 (1054.0/763.0)
|  |  DISTTOWK > 6.1105
|  |  |  TOUR_FLG <= 0
|  |  |  |  STOPS <= 0
|  |  |  |  |  BEGNTIME <= 844: A3 (10754.0/2317.0)
|  |  |  |  |  BEGNTIME > 844
|  |  |  |  |  |  CNTTDTR <= 2: A3 (6078.0/1856.0)
|  |  |  |  |  |  CNTTDTR > 2
|  |  |  |  |  |  |  Unique_Persons.HBRESDN <= 50: A3 (6250.0/3711.0)
|  |  |  |  |  |  |  Unique_Persons.HBRESDN > 50
|  |  |  |  |  |  |  |  Unique_Persons.TRAVDAY <= 1: A1 (3764.0/2466.0)
|  |  |  |  |  |  |  |  Unique_Persons.TRAVDAY > 1
|  |  |  |  |  |  |  |  |  Unique_Persons.TRAVDAY <= 6
|  |  |  |  |  |  |  |  |  |  BEGNTIME <= 1822
|  |  |  |  |  |  |  |  |  |  |  BEGNTIME <= 1046: A3 (1771.0/955.0)
|  |  |  |  |  |  |  |  |  |  |  BEGNTIME > 1046
|  |  |  |  |  |  |  |  |  |  |  |  ENDTTIME <= 1314: A1 (2659.0/1583.0)
|  |  |  |  |  |  |  |  |  |  |  |  ENDTTIME > 1314: A3 (9084.0/4806.0)
|  |  |  |  |  |  |  |  |  |  BEGNTIME > 1822
|  |  |  |  |  |  |  |  |  |  |  ENDTTIME <= 2019: A1 (2470.0/1560.0)
|  |  |  |  |  |  |  |  |  |  |  ENDTTIME > 2019: A3 (2087.0/1416.0)
|  |  |  |  |  |  |  |  Unique_Persons.TRAVDAY > 6: A1 (4186.0/2745.0)
|  |  |  |  STOPS > 0: A1 (6148.0/3163.0)
|  |  |  TOUR_FLG > 0
|  |  |  |  STOPS <= 2
|  |  |  |  |  USEINTST <= 1: B (4581.0/3084.0)
```

```
|   |   |   |   |   USEINTST > 1
|   |   |   |   |   |   DISTTOWK <= 8.3325: A3 (3867.0/2310.0)
|   |   |   |   |   |   DISTTOWK > 8.3325: B (2946.0/1883.0)
|   |   |   |   STOPS > 2: B (1317.0/847.0)
DISTTOWK > 10.5545
|   DISTTOWK <= 20
|   |   CNTTDTR <= 2: B (14432.0/3526.0)
|   |   CNTTDTR > 2
|   |   |   TOUR_FLG <= 0
|   |   |   |   STOPS <= 0
|   |   |   |   |   ENDTTIME <= 1424
|   |   |   |   |   |   BEGNTIME <= 946: B (13104.0/4169.0)
|   |   |   |   |   |   BEGNTIME > 946
|   |   |   |   |   |   |   USEINTST <= 1
|   |   |   |   |   |   |   |   CNTTDTR <= 4: B (1563.0/1012.0)
|   |   |   |   |   |   |   |   CNTTDTR > 4: A1 (4151.0/2715.0)
|   |   |   |   |   |   |   USEINTST > 1: A1 (6492.0/3857.0)
|   |   |   |   |   ENDTTIME > 1424
|   |   |   |   |   |   BEGNTIME <= 1743
|   |   |   |   |   |   |   ENDTTIME <= 1751
|   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY <= 1: A1 (1674.0/1186.0)
|   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY > 1
|   |   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY <= 6: B (9176.0/4298.0)
|   |   |   |   |   |   |   |   |   Unique_Persons.TRAVDAY > 6: A1 (1833.0/1320.0)
|   |   |   |   |   |   |   ENDTTIME > 1751: B (1200.0/466.0)
|   |   |   |   |   |   BEGNTIME > 1743
|   |   |   |   |   |   |   ENDTTIME <= 1815: A1 (1152.0/631.0)
|   |   |   |   |   |   |   ENDTTIME > 1815
|   |   |   |   |   |   |   |   BEGNTIME <= 1810: B (1061.0/413.0)
|   |   |   |   |   |   |   |   BEGNTIME > 1810
|   |   |   |   |   |   |   |   |   ENDTTIME <= 2127
|   |   |   |   |   |   |   |   |   |   ENDTTIME <= 1846: A1 (1181.0/650.0)
|   |   |   |   |   |   |   |   |   |   ENDTTIME > 1846
|   |   |   |   |   |   |   |   |   |   |   BEGNTIME <= 1845: B (1087.0/568.0)
|   |   |   |   |   |   |   |   |   |   BEGNTIME > 1845
|   |   |   |   |   |   |   |   |   |   |   |   ENDTTIME <= 1926: A1 (1155.0/642.0)
|   |   |   |   |   |   |   |   |   |   |   ENDTTIME > 1926
|   |   |   |   |   |   |   |   |   |   |   |   |   Unique_Persons.HTPPOPDN <= 300: B (1604.0/1029.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   Unique_Persons.HTPPOPDN > 300: A1 (3084.0/2033.0)
|   |   |   |   |   |   |   |   |   ENDTTIME > 2127: B (2269.0/1367.0)
|   |   |   |   STOPS > 0: A1 (7848.0/3991.0)
|   |   |   TOUR_FLG > 0: B (16431.0/9524.0)
```

```
|   DISTTOWK > 20
|   |   DISTTOWK <= 30
|   |   |   CNTTDTR <= 2: C (6599.0/1749.0)
|   |   |   CNTTDTR > 2
|   |   |   |   TOUR_FLG <= 0
|   |   |   |   |   STOPS <= 0
|   |   |   |   |   |   Unique_Persons.TRAVDAY <= 1: A1 (2759.0/1898.0)
|   |   |   |   |   |   Unique_Persons.TRAVDAY > 1
|   |   |   |   |   |   |   Unique_Persons.TRAVDAY <= 6
|   |   |   |   |   |   |   |   BEGNTIME <= 1811
|   |   |   |   |   |   |   |   |   ENDTTIME <= 1429
|   |   |   |   |   |   |   |   |   |   BEGNTIME <= 1053: C (4801.0/1458.0)
|   |   |   |   |   |   |   |   |   |   BEGNTIME > 1053: A1 (2283.0/1412.0)
|   |   |   |   |   |   |   |   |   ENDTTIME > 1429: C (4754.0/2505.0)
|   |   |   |   |   |   |   |   BEGNTIME > 1811: A1 (3078.0/2251.0)
|   |   |   |   |   |   |   Unique_Persons.TRAVDAY > 6: A1 (3128.0/2274.0)
|   |   |   |   |   STOPS > 0: A1 (3230.0/1704.0)
|   |   |   |   TOUR_FLG > 0: C (6725.0/4386.0)
|   |   DISTTOWK > 30
|   |   |   DISTTOWK <= 40
|   |   |   |   BEGNTIME <= 745: D (3021.0/763.0)
|   |   |   |   BEGNTIME > 745
|   |   |   |   |   CNTTDTR <= 3: D (2985.0/1448.0)
|   |   |   |   |   CNTTDTR > 3
|   |   |   |   |   |   TOUR_FLG <= 0
|   |   |   |   |   |   |   STOPS <= 0
|   |   |   |   |   |   |   |   USEINTST <= 1
|   |   |   |   |   |   |   |   |   CNTTDTR <= 4: D (1085.0/635.0)
|   |   |   |   |   |   |   |   |   CNTTDTR > 4
|   |   |   |   |   |   |   |   |   |   ENDTTIME <= 1529: A1 (1001.0/687.0)
|   |   |   |   |   |   |   |   |   |   ENDTTIME > 1529: D (1402.0/1064.0)
|   |   |   |   |   |   |   |   USEINTST > 1: A1 (2697.0/1742.0)
|   |   |   |   |   |   |   STOPS > 0: A1 (1209.0/629.0)
|   |   |   |   |   |   TOUR_FLG > 0
|   |   |   |   |   |   |   ENDTTIME <= 1531: B (1005.0/790.0)
|   |   |   |   |   |   |   ENDTTIME > 1531: D (1067.0/761.0)
|   |   |   DISTTOWK > 40
|   |   |   |   DISTTOWK <= 50
|   |   |   |   |   BEGNTIME <= 746: E (1371.0/361.0)
|   |   |   |   |   BEGNTIME > 746
|   |   |   |   |   |   CNTTDTR <= 3: E (1425.0/713.0)
|   |   |   |   |   |   CNTTDTR > 3
```

```
|   |   |   |   |   |   |   USEINTST <= 1
|   |   |   |   |   |   |   |   CNTTDTR <= 5: E (1036.0/747.0)
|   |   |   |   |   |   |   |   CNTTDTR > 5: A1 (1191.0/901.0)
|   |   |   |   |   |   |   USEINTST > 1: A1 (1854.0/1289.0)
|   |   |   |   DISTTOWK > 50
|   |   |   |   |   DISTTOWK <= 60
|   |   |   |   |   |   CNTTDTR <= 4: F (1567.0/751.0)
|   |   |   |   |   |   CNTTDTR > 4: A1 (1678.0/1274.0)
|   |   |   |   |   DISTTOWK > 60
|   |   |   |   |   |   DISTTOWK <= 80
|   |   |   |   |   |   |   DISTTOWK <= 69: G (1411.0/947.0)
|   |   |   |   |   |   |   DISTTOWK > 69: H (1201.0/941.0)
|   |   |   |   |   |   DISTTOWK > 80: A1 (2089.0/1623.0)
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Identification of common characteristics in activity-travel behavior based on activity-travel diaries

Richting: **Master of Transportation Sciences-Mobility Management**
Jaar: **2015**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Strackx, Nick

Datum: **28/08/2015**