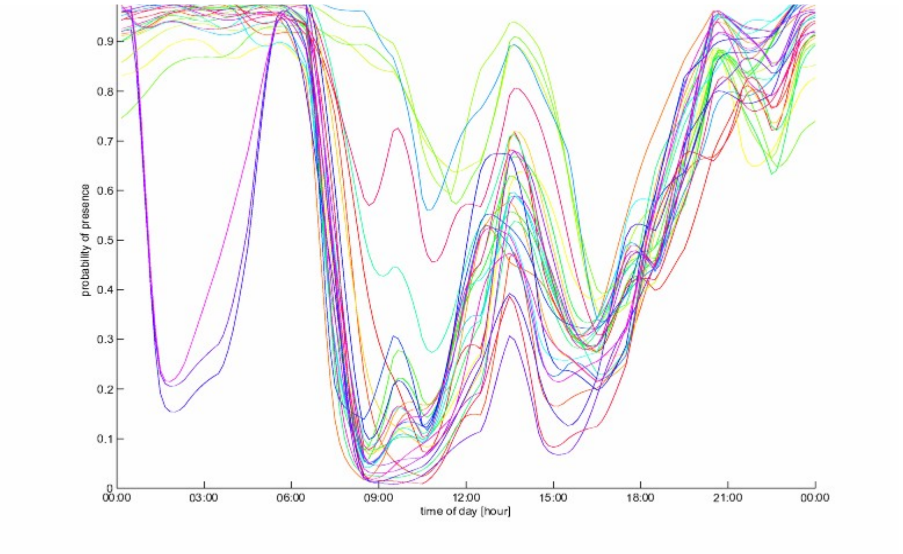
# Bottom-up modelling for stochastic prediction of residential and work-place occupancy



Grégoire Virard
Master in Energy Management and Sustainability

Semester project
June 2012

Directed by :
Urs Wilke
LESO-PB : Sustainable Urban Development group

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

# Table of content

1Abstract....3
2Introduction....3
3Methodology....3
3.1The survey....3
3.2Model ....4
3.3Theory ....6
3.3.1Markov chain and Master equation....6
3.3.2Logistic regression....7
3.3.3Interval's and data processing....7
3.4Two models....8
3.4.1Simple model....8
3.4.2Detailed model....10
4Modelling Results....13
4.1Population distribution of presence probabilities....13
4.2Results....15
4.2.1Household occupancy....15
4.2.2Work-place occupancy....18
4.3Discussion....21
4.4Application....21
4.5Outlook....21
5Conclusion....22
6References....23
7Appendix....24

# Index of figures

Figure 1: scheme for transition probabilities and Master equation....6
Figure 2: Scheme for the interval's processing....8
Figure 3: The 4 sub-populations....8
Figure 4: Estimated parameters for the "coming" case....9
Figure 5: Estimated parameters for the "leaving" case....9
Figure 6: Presence profiles, results of the simple model....10
Figure 7: Example of results from the Biogeme estimation of utility parameters....12
Figure 8: Profile generation process....13
Figure 9: 30 individual household presence profiles randomly chosen over the whole population. .14
Figure 10: presence probabilities profile for the whole population....15
Figure 11: presence probabilities profile for a week day for the active population ....16
Figure 12: presence probabilities profile for a week day for the retired population....16
Figure 13: presence probabilities profile for men....17
Figure 14: presence probabilities profile for women....17
Figure 15:  presence probabilities profile for the whole population....18
Figure 16: presence probabilities profile for a week day for the active population....19
Figure 17: presence probabilities profile for a week day for the retired population....19
Figure 18: presence probabilities profile for men....20
Figure 19: presence probabilities profile for women....20

# 1 Abstract

This study presents the establishment of a model that predicts the occupancy in dependence of time of day in the residential or work-place contexts as a function of the individual characteristics of the population. It is a stochastic approach based on a Markov chain and thus the master equation can be used to describe the dynamical behaviour. The presence probabilities have been derived from the transition probabilities (e.g. leaving/coming probabilities) calibrated based on a large time-use survey using logistic regression as a function of individual characteristics. The project has shown good results for the prediction of overall tendencies of presence profiles and has already given the track to give reliable and accurate results with some improvements.

# 2 Introduction

Currently, no time-dependent model exists for this purpose where specific human characteristics are included while it has many potential applications in different research fields, such as, sociology and economy and in particular building physics, where the human actions are decisive (e.g. manual control of windows or blinds of a house façade). This is especially relevant in the topic of energy savings and optimization in the actual context of climate change and energy crisis where the model resulting from this project could be used in the prediction of needs for heating, cooling, lighting, etc...

To go through the development of this model, the methodology used will firstly be presented with a brief description of the concerned time-use survey. Next, the concepts and the theory behind the present project are introduced. It is presented the Markov process and master equation that are used to calculate the time-dependent presence probabilities that are derived from the transition probabilities calibrated with a binary logistic regression. Then the model will be presented more into details, in particular with the usage of  a z-test in order to assess the significance of the parameters involved. After all comes the presentation of the final results of this study in terms of occupancy profiles in households and work-places. Finally, an example of its application is proposed for the thermal simulation of a building that takes into account the heat gains of the occupants.

# 3 Methodology

In this section, the survey the present project is based on is firstly introduced, then, the basic concepts and assumptions for the implementation of the model are presented. Next, a section introduces the theoretical elements used in the approach. Finally, two levels of complexity of the models are presented: a simple one to illustrate/test the methodology and a detailed one. The methodology is used to model home occupancy as well as work-place occupancy.

## *3.1 The survey*

The development of the model presented in this study is based on a time-use survey (TUS) done in France by the National Institute of Statistics and Economic Studies from February 16th 1998 to February 14th 1999[1]. 15441 people (14631 adults & 810 children) completed questionnaires describing the chronological course of activities in their diaries in a resolution of 10 min time-steps during 24 h starting and ending at midnight. Those surveyed were asked to specify the activities they performed, as well as the location (at own home and work-place), and the corresponding time of day by filling out a questionnaire[2]. Other types of location such as fitness centre or restaurant were derived from the corresponding activity types recorded. The following diary is given as an

[1] htpp://www.insee.fr

[2] K. Fisher et al., October 2011, Multinational Time Use Study - User’s Guide and Documentation

example :

- 0:00am to 8:00am : at home
- 8:00am to 8:30 am : travelling
- 8:30am to 12:00am : at work-place
- 12:00am to 1:00pm : at restaurant
- 1:00pm to 5:00pm : at work-place
- 5:00pm to 5:30pm : travelling
- 5:30pm to 6:30pm : sport training
- 6:30pm to 6:50pm : travelling
- 6:50pm to 0:00pm : at home

Furthermore, a wide range of individual details and characteristics was recorded such as age, gender, education level, income, marital status, number of children, etc... In total, there are 139 variables for each surveyed person grouped in five domains : diary informations, household-level variables, person-level variables, employment & education level and health variables [*2*].

*Note : a complete description is included and available in electronic format in a database containing TUS data of several countries and years at* http://www.timeuse.org.

The present study focuses on the household and work-place occupancies, therefore, the attention is put on the variables of presence (at home and at work-place) and individual and household characteristics that will be detailed in section 3.4.2. The detailed contextual recordings in this survey are very appropriate with regard to the stochastic approach of the project.

## 3.2 Model

The main quantity of interest in this research study is the probability to be present in the type of location (home/work) in dependence of time of day and individual characteristics. The model that is used is based on a non-homogeneous Markov process, which will be described more into details in *section 3.3*. Therefore, the individuals are assumed to behave independently (necessary condition of Markov processes), meaning that the presence of one individual is supposed to be independent from the presence state of any other person. Although the presence as a function of time of day is in general a continuous function, it can be well approximated as a discretized time series, that is necessary regarding the discretized nature of the monitored data (*cf. section 3.1*). For convenience, the resolution of the time-use survey of 10 min. is chosen as the length of one time step. The presence probability at a time step $t_i$ is supposed to only be dependent of the preceding presence at the time $t_{i-1.}$ The driving forces in this Markov process are represented by time-dependent transition probabilities *(cf. fig.1 & section 3.3.1*) that are calculated using utility functions within a logistic regression *(cf. section 3.3.2*). The dynamical behaviour of a system with the Markov property is governed by the Master equation which also will be detailed in the next section. Consequently, a large part of this project consists in estimating the transition probabilities in order to derive the presence probabilities.

To characterize the household occupancy as a function of individual characteristics, all the relevant variables have to be carefully defined in a way they are easily expressible Here the variables were expressed as dummy variables (can only take the value 1 or 0), as in general, one can not assume a monotonic increase/decrease of the transition probabilities with increasing value of a given continuous predictor/independent variable (see definition of utility functions in *section 3.3.2* . However, some TUS variables can take more than two values (e.g. : the variable representing the income range can take 10 different values representing the levels of income and other possibilities such as “doesn't know”), These were translated into multiple dummy variables (e.g. : for the income level, it yields to 5 dummy variables : 1) earning less than 700€, 2) between 700€ and 1500€, 3) between 1500€ and 3000€, 4) more than 3000€, or 5) “doesn't know”).

However, please notice that the assumptions made in this project in order to use a Markov process do not necessarily reflect all the realistic dependencies regarding the probabilities of presence. In fact, a Markov chain does not take into account the influence of other people on the presence probability of someone (e.g. influence of the presence of one member of a couple on the presence of the other one). Furthermore the Markov process used in the present study cannot reflect the influence of the presence at a time-step earlier than the precedent step, this would need a higher order Markov chain. And a Markov chain is not able to take into account a state that will occur later (e.g. "I am at my office, it is 3:00pm and I know that I will leave the work-place at 6:30pm", the probability to leave the work-place at the next time-step would be different than if "I do not know at what time I will leave my office today").

## *3.3 Theory*

### 3.3.1 Markov chain and Master equation

In a first-order Markov chain, the next state ( $X_{n+1}$) does not depend on the sequence of precedent states, but only on the current state ( $X_{n+1}$) (see *eq. 1*). This is called as a memoryless random process.

$$Pr(X_{n+1}=x\,|X_1=x_1,X_2=x_2,\ldots,X_n=x_n) = Pr(X_{n+1}=x\,|X_n=x_n) \qquad (eq.\ 1)$$

In the present study, the presence probability at a time-step is governed by the transition probabilities as well as the value of the preceding time step as described in *fig. 1* with the following explanations.

- 0 and 1 correspond to "away" and "at home" states respectively
- The transition probabilities are given by:
  - $t_{0\to0}$ : probability of "staying away"
  - $t_{0\to1}$ : probability of "coming at home"
  - $t_{1\to1}$ : probability of "staying at home"
  - $t_{1\to0}$ : probability of "leaving home"

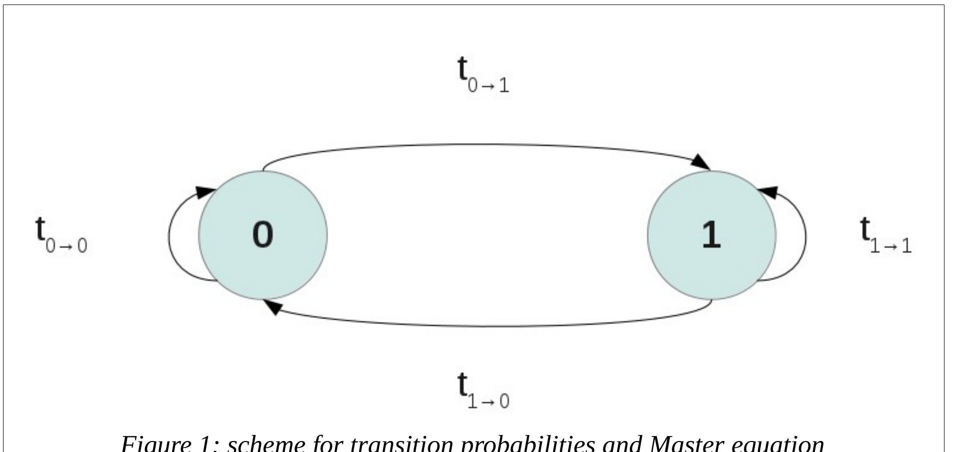


*Figure 1: scheme for transition probabilities and Master equation*

The transition probabilities are grouped in the transition matrix T (see *eq. 2*) which is used in the Master equation (*eq. 3*) that governs the dynamics between each time-step.

$$T=\begin{pmatrix} t_{0\to0} & t_{1\to0} \\ t_{0\to1} & t_{1\to1}\end{pmatrix} \quad ; \quad \text{and} \quad \begin{Bmatrix} t_{0\to1}+t_{0\to0}=1 \\ t_{1\to1}+t_{1\to0}=1 \end{Bmatrix} \qquad (eq.\ 2)$$

$$\frac{d\vec{P}}{dt}=T(t)\vec{P} \quad \text{; where P=(P}_0\text{,P}_1\text{) is the vector of presence probabilities} \qquad (eq.\ 3)$$

By rewriting as a discrete process between the next and the actual time-steps, it yields :

$$\frac{P(t_{n+1})-P(t_n)}{t_{n+1}-t_n}=T(t_n)*P(t_n) \qquad (eq.\ 4)$$

And

$$P(t_{n+1})=P(t_n)*\left[1+(t_{n+1}-t_n)*T(t_n)\right] \qquad (eq.\ 5)$$

In the logistic regression which will be presented in the next section the quantities that calibrated express rather a transition rate than a probability. Therefore, in eq. 5 the term $(t_{n+1}-t_n)*T(t_n)$ corresponds to a matrix of transition probabilities for the corresponding time-interval. Thus, it follows for the first component of the two-dimensional presence vector (probability calculation of being present at $t_{n+1)}$:

$$P_1(t_{n+1})=P_1(t_n)*\left(1-t_{1\rightarrow 0}(t_n)\right)+P_0(t_n)*t_{0\rightarrow 1}(t_n) \qquad (eq.\ 6)$$

This means that the presence probability at the next time step can be calculated by decreasing it by the “share” that someone is leaving $P_1(t_n)*\left(1-t_{1\rightarrow 0}(t_n)\right)$ and by increasing it by the “share” that someone is arriving $P_0(t_n)*t_{0\rightarrow 1}(t_n)$ .

### 3.3.2 Logistic regression

In order to express the transition probabilities presented above as a function of the characteristics of the population and time of day, a binomial logit model (also called binary logistic regression) is used, in which the deterministic parts of the transition probabilities depend on utility functions. The approach is detailed in this section.

First of all, the utility functions are defined as linear combinations of *n* dummy variables (note : the number of parameter *n* will be discussed later on *(cf. section 3.4.2))* that represent the individual's characteristics (see *eq.* 7). The set ($V_0$, $V_1$) of utility functions is used in each transition probability estimation (cf. *eq. 8*) where : the $x_i$ are the dummy variables that express the characteristics of the population as discussed in the *section 3.2*, the parameters $\beta_i$ are constants for each variable that have to be calibrated at each time-step, and $\alpha_0$, $\alpha_1$ are alternative-specific constants (ASCs) for each utility function. Please note that $V_0$ is always fixed without loss of generality, in fact only the difference of both utility functions affects the probabilities which have to be estimated (*cf.* eq. 8)

$$\begin{Bmatrix} V_0(x_1,\ldots,x_n,t)=\alpha_0+\beta_{0,1,t}\cdot x_1+\beta_{0,2,t}\cdot x_2+\ldots+\beta_{0,n,t}\cdot x_n=0 \quad \text{fixed} \\ V_1(x_1,\ldots,x_n,t)=\alpha_1+\beta_{1,1,t}\cdot x_1+\beta_{1,2,t}\cdot x_2+\ldots+\beta_{1,n,t}\cdot x_n \end{Bmatrix} \qquad (eq.\ 7)$$

The parameters (α and $\beta_i$) of the utility functions are estimated at each time-step with the help of the software Biogeme [3,5]. Biogeme is a tool created by Prof. Michel Bierlaire (EPFL) initially developed for transport optimisation application. In general, it is designed for research in the context of discrete choice models and therefore appropriate for this project.

Each transition probability is calibrated by estimating the parameter values that maximize the loglikelihood of the model. Each logit model ($V_0$, $V_1$) estimated allows to calibrate two transition probabilities (of the four matrix entries), one and its complementary (*cf. eq.* 2, e.g. “leaving home” probability and “staying at home” probability, which always sum up to one). So at each time-step, two logit model are estimated to compute the whole set of transition probabilities.

$$t_{transition}(x_1,\ldots,x_n,t)=\frac{1}{1+\exp\left(V_0(x_1,\ldots,x_n,t)-V_1(x_1,\ldots,x_n,t)\right)} \qquad (eq.\ 8)$$

### 3.3.3 Interval's and data processing

As the transition probabilities depend on the time of day, they are estimated for each hour interval of the day. In order to use the data from the survey, it is necessary to interpret and adapt them for the specific purpose of this study. As explained in section 3.1, the data contains activity states in a resolution of 10 min. The first step is to extract the binary presence state information of the considered type of place (work/home) and to remove the activity information, which is of no use in this topic. This means that each time the presence state changes (from “present” to “away” and vice-versa), we get the new state of presence (“present” or “away”), the time of the beginning of the state and the end time of the state. The calibration of the logistic regression models has to be based on the events in the TUS data that are overlapping with the hour interval considered. There are two types of events, “coming” or “leaving”. As mentioned before (*cf. section 3.3.2*), complementary transition probabilities allow to consider only these two events in order to calibrate their transition probabilities. In this regard, the corresponding events have to be extracted out from the entire set of events. This is illustrated in the figure below (identical procedures for both home and work-place occupancies).

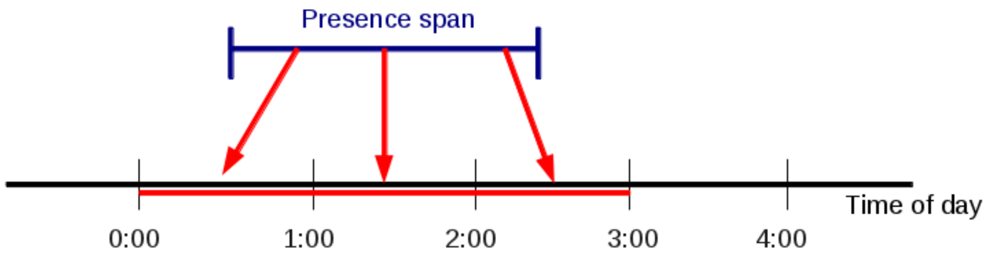


*Figure 2: Scheme for the interval's processing*

Then, the utility functions are calibrated on the basis of the TUS data. The data are split into 48 files with two files for each hour of day, one for the “coming” state and one for the “leaving” state. And so, two hourly utility functions are calibrated, for the “coming” and “leaving” transition probabilities respectively. This splitting option has been undertaken in order to restrict the set of events that is used to calibrate the transition probabilities to the relevant ones (e.g., in *fig. 2* the presence state, schematized by the blue line, is of relevance for the three red intervals) and to reduce the data processing load of Biogeme. For each hour, Biogeme reads the corresponding data file and the values of the parameters ($\alpha$ and $\beta_i$) of the utility functions are optimized.

Now all the needed tools are introduced, allowing to predict occupancy as a function of time of day and individual characteristics.

## 3.4 Two models

The TUS data contain 139 individual variables, which can often take more than two values. This means that there are more than $2^{139}=7*10^{41}$ possible combinations. As the model is only calibrated on a set of about 15'000 individuals this demonstrates the fact that it is meaningless to calibrate a model which depends on the complete set of individual characteristics. First the approach is tested with a very simple model which also illustrates of the methodology to clarify the procedure.

### 3.4.1 Simple model

In order to test the methodology, the simple model to begin with only takes into account two parameters : $x_1$ : gender (man or woman), and $x_2$ : age (≤ 65 years old or > 65 years old). These two parameters subdivide the population into $2^2=4$ sub-populations depending on the value of the vector X=( $x_1$, $x_2$) described hereafter :

- (1,1) : man & > 65
- (1,0) : man & ≤ 65
- (0,1) : woman & > 65
- (0,0) : woman & ≤ 65

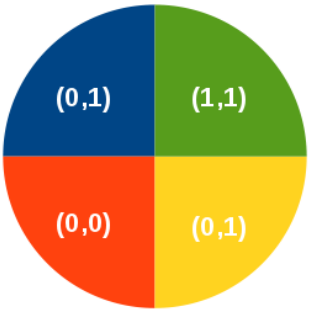


*Figure 3: The 4 sub-populations*

This corresponds to the following utility functions :

$$\begin{cases} V_0(x_1,x_2,t)=\alpha_0+\beta_{0,gender,t}\cdot x_1+\beta_{0,age,t}\cdot x_2=0 \quad \text{fixed} \\ V_1(x_1,x_2,t)=\alpha_1+\beta_{1,gender,t}\cdot x_1+\beta_{1,age,t}\cdot x_2 \end{cases} \qquad (eq.\ 9)$$

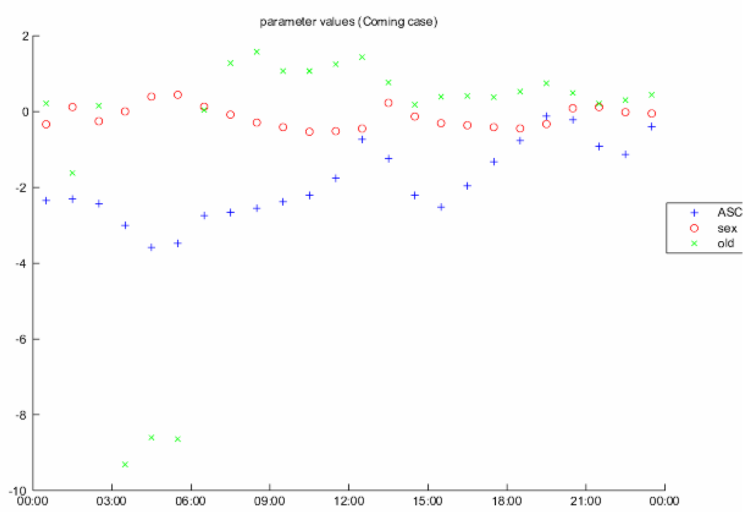


*Figure 4: Estimated parameters for the "coming" case*

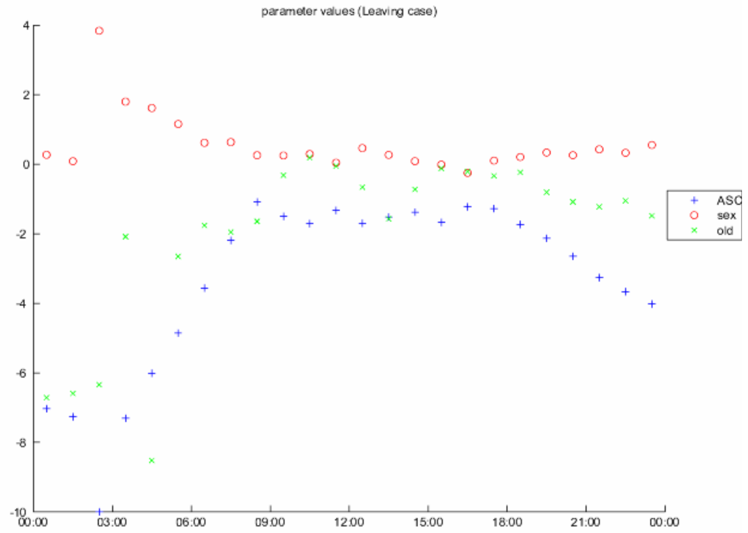


*Figure 5: Estimated parameters for the "leaving" case*

The parameters of the 48 utility functions ($V_1$) are estimated using the logistic regression approach (*cf*. section 3.3.2 & 3.3.3). In *Figs. 4 & 5*, the values of these parameters are illustrated for the “coming” and “leaving” transitions, respectively. The values of the parameter describing the age are represented in green, in red the gender parameter, and in blue the alternative specific constants (ASCs) that represent the utility constant (α). Notice that only the difference between the two utility functions ($V_1$-$V_0$) affects the choice probability (cf. Eq. 8). Consequently, all parameters in $V_0$ were fixed to zero. In *fig.* 4 (coming case) are represented the values of the parameters of the utility function $V_1$ that describes the “coming” utility ($V_0$ corresponds to “staying away”), and *fig. 5* is for the “leaving” utility $V_1$ parameters ($V_0$ corresponds to “staying at home”).The corresponding transition probabilities are calculated using eq. 8. The time-dependent presence probabilities are then calculated by incrementally applying the Master equation. Here, the initial value (at time 0:00am) is set to the according value in the TUS data. f*ig. 6* shows a comparison of the presence probabilities of the four sub-populations derived from the model (full lines) against the profiles that were observed in the raw TUS data (bullets). The latter were calculated, for each sub-population and at each time step, by the ratio of the accumulated presence of all people and the number of individuals in the whole sub-population. In green is represented the “women older than 64 years” sub-population, in blue the “men older than 64 years” sub-population, in black the “women younger than 64 years” sub-population, and in red the “men younger than 64 years” sub-population.

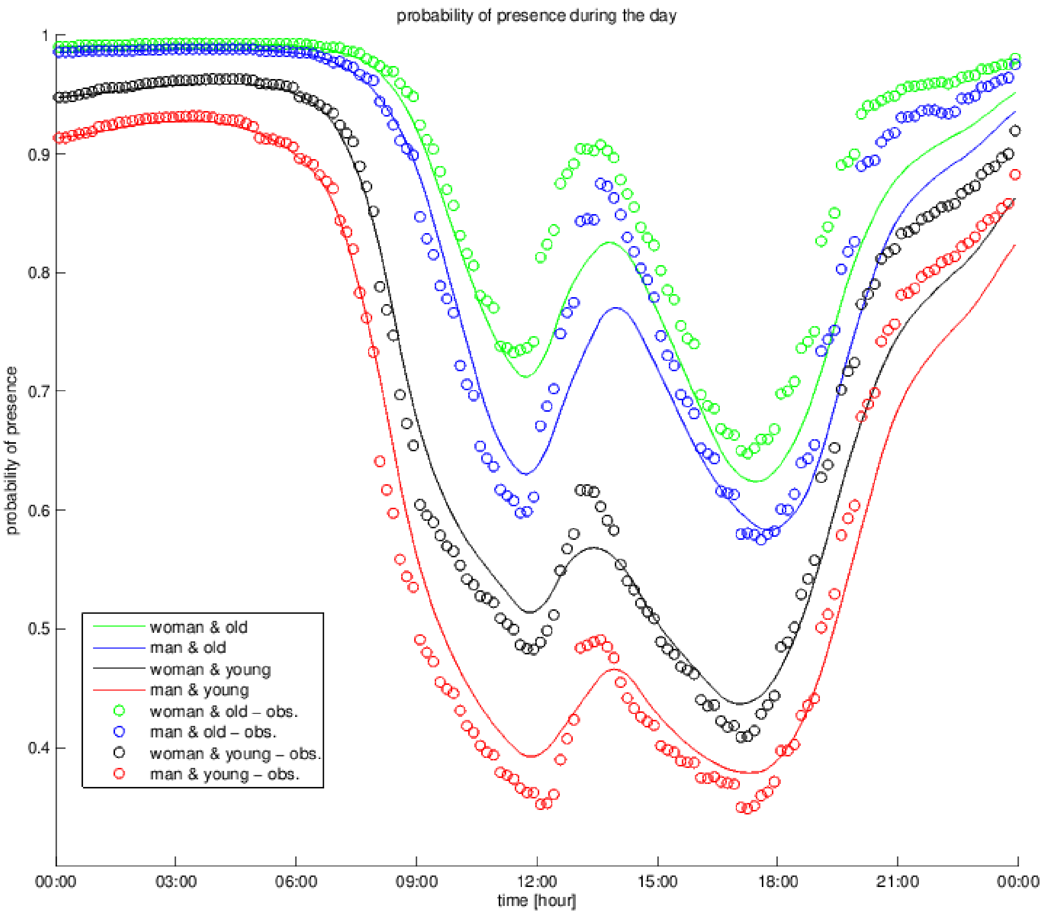


*Figure 6: Presence profiles, results of the simple model*

### 3.4.2 Detailed model

After testing the approach with the simple model presented before, it will now be adapted to a more complex one. Due to multicollinearity reasons, it would not be feasible to estimate the transition probabilities when the whole set of dummy variables (*cf*. list below) is included in the utility functions. Therefore, the main challenge here consists in choosing the parameters to include in the model with the aim of preparing a parsimonious model without influences of insignificant parameters, which is at the same time as detailed as possible to yield a high level of predictivity.

In a first step, the number of parameters is limited to 18 (out of the initially 139 ones) considered as possibly relevant. These 18 parameters, detailed hereafter represent finally 54 linearly independent dummy variables.

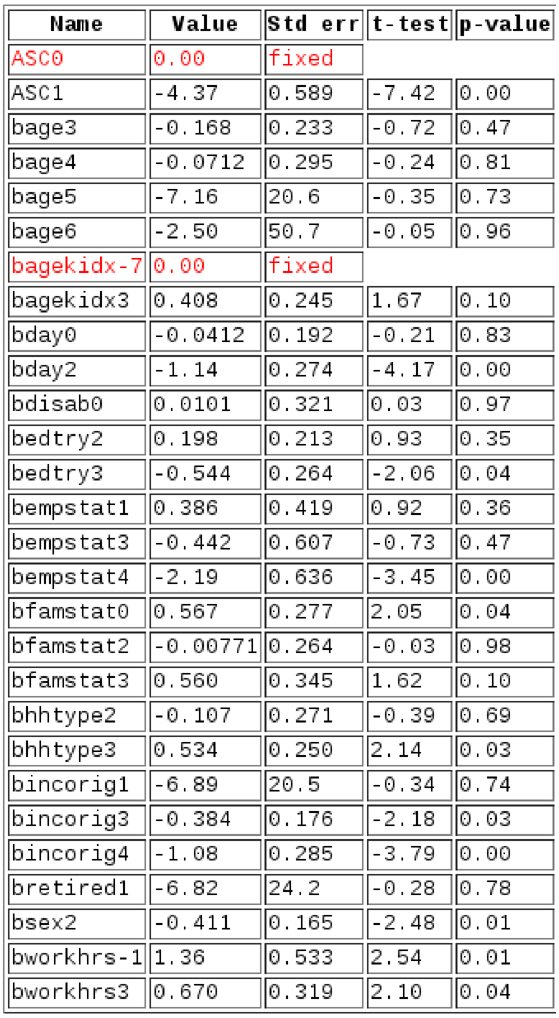
List and description of arbitrary chosen parameters[3]:

- day of the week (Monday to Thursday, Friday, or weekend)
- gender (male/female)
- age (<18, 18-35, 36-45, 46-60, 61-75, >75)
- type of household (one person household, couple, …)
- age of the youngest kid in the household (0-4, 5-12, 13-17, ≥18, “no child in the household”)
- income level (monthly) (<700€, 700€ - 1500€, 1500€ - 3000€, > 3000€, or “doesn't know”)
- own home (own, rent, other)
- type of area : urban or rural
- family status : presence of children in the household and range of ages
- being a single parent or not
- civil status (being in couple or not)
- employment status (full time, part-time, employed but unknown status, “not in paid work”)
- being a student or not
- number of weekly working hours (<16, 16-35, 36-45, >45, “not asked or not answered”)
- level of education (uncompleted, secondary completed, above)
- taking routinely care about someone with disability or health concern in the household
- having a disability or a long-term limiting health condition

After the pre-selection of these dummy variables, an automatized algorithm was used, to only include variables in the utility functions, where there is a significant difference in behaviour between the two sub-populations classified by the two values the dummy variables can take. This was done by using a two proportion z-test (proportion of people leaving/coming of all people at home/not at home). Only the variables $x_i$ were put into the model, where the corresponding sub-populations ( $x_i$=0 or $x_i$=1) have a significantly different behaviour with respect to the two proportion z-test. Note : the two-proportion z-test is only reliable for sub-populations with more than 30 elements, and thus, variables leading to sub-populations smaller than 30 elements are systematically omitted in the model.

[3] From or adapted from [2,4]

Furthermore, a third consideration has to be undertaken concerning some multicollinearities or even linear dependencies between (a set of) variables that arise during the estimation and may cause the diverging standard deviations of the corresponding parameters, even after the two precedent steps. It was observed that some variables are dependent (e.g. : the dummy variable for the family status that represents a household with a child living with his parents, and the variable for the age of the youngest kid in the household are obviously linear dependent). Consequently, the values of some of the concerned variables have been fixed to zero in order to erase these dependencies, correct and reframe the model.



| Name | Value | Std err | t-test | p-value |
|---|---|---|---|---|
| ASC0 | 0.00 | fixed | | |
| ASC1 | -4.37 | 0.589 | -7.42 | 0.00 |
| bage3 | -0.168 | 0.233 | -0.72 | 0.47 |
| bage4 | -0.0712 | 0.295 | -0.24 | 0.81 |
| bage5 | -7.16 | 20.6 | -0.35 | 0.73 |
| bage6 | -2.50 | 50.7 | -0.05 | 0.96 |
| bagekidx-7 | 0.00 | fixed | | |
| bagekidx3 | 0.408 | 0.245 | 1.67 | 0.10 |
| bday0 | -0.0412 | 0.192 | -0.21 | 0.83 |
| bday2 | -1.14 | 0.274 | -4.17 | 0.00 |
| bdisab0 | 0.0101 | 0.321 | 0.03 | 0.97 |
| bedtry2 | 0.198 | 0.213 | 0.93 | 0.35 |
| bedtry3 | -0.544 | 0.264 | -2.06 | 0.04 |
| bempstat1 | 0.386 | 0.419 | 0.92 | 0.36 |
| bempstat3 | -0.442 | 0.607 | -0.73 | 0.47 |
| bempstat4 | -2.19 | 0.636 | -3.45 | 0.00 |
| bfamstat0 | 0.567 | 0.277 | 2.05 | 0.04 |
| bfamstat2 | -0.00771 | 0.264 | -0.03 | 0.98 |
| bfamstat3 | 0.560 | 0.345 | 1.62 | 0.10 |
| bhhtype2 | -0.107 | 0.271 | -0.39 | 0.69 |
| bhhtype3 | 0.534 | 0.250 | 2.14 | 0.03 |
| bincorig1 | -6.89 | 20.5 | -0.34 | 0.74 |
| bincorig3 | -0.384 | 0.176 | -2.18 | 0.03 |
| bincorig4 | -1.08 | 0.285 | -3.79 | 0.00 |
| bretired1 | -6.82 | 24.2 | -0.28 | 0.78 |
| bsex2 | -0.411 | 0.165 | -2.48 | 0.01 |
| bworkhrs-1 | 1.36 | 0.533 | 2.54 | 0.01 |
| bworkhrs3 | 0.670 | 0.319 | 2.10 | 0.04 |

*Figure 7: Example of results from the Biogeme estimation of utility parameters*

*Fig. 7* shows one of the 48 results for the estimation of utility function parameters with Biogeme. Showing the parameter values for the probability of “coming to work between 7:00am and 8:00am”. The meaning of the parameter names from the table is shown in *table 1* in appendix. See the different values and standard deviations given for each parameter, and the parameter manually fixed to zero (“age of the youngest kid in the household”). The row “Value” gives the resulting estimation of the parameters. e.g. *bempstat4* (“not in a paid work” parameter) being negative, means that for someone who is not employed in a paid work, the probability to come at work between 7:00am and 8:00am is decreased. Additionally, Std err” shows the standard error of the estimated parameter, and the t test shows the ratio of Value/Std err. The according “p-value” row describes the significance of each parameter in the model (0≤ p-value ≤1), i.e. the two proportion z-test gives the significance of a parameter isolated (only between the two sub-populations created 1/0) but here the significance is estimated over all the parameters expressed in the model (resulting from the z-test) with eventual correlations between the parameters. The lower the p-value is, the more significant is the corresponding parameter. However, one can notice that several parameters have high p-values meaning a low significance of these parameters in the model. These non-significant parameters create some variations and non-accuracies that are found in *fig. 9* with kinks or breaks that are detailed later on.

# 4 Modelling Results

## 4.1 Population distribution of presence probabilities

The presence probabilities are calculated for all the approx. 15'000 peoples in order to compute individual presence profiles. It is assumed that (sub-)population distribution of presence probabilities can be specified by sample distribution of presence profiles of all individuals contained in the sample. In f*ig. 8* is shown a sample of 30 individual profiles randomly chosen over the whole population (~15'000 persons). The initial presence probability values, requested for the incremental process of the Markov chain with the Master equation, have been computed as follow : the diaries are 24h based and the probability at 0:00am should be the same as at 0:00pm, so the presence probabilities calculation has been repeated over a longer period (4*24h) and only the last 24h were recorded. For this process the first probability was set at 0.98 and 0.10 for “at home” and “at work-place” respectively. *Fig. 7* which illustrates the calibration process for one profile randomly chosen. One can observes that after the second step of 24h, the profile is already stabilised (see the red circles that emphasise the adjustment of the initial value).

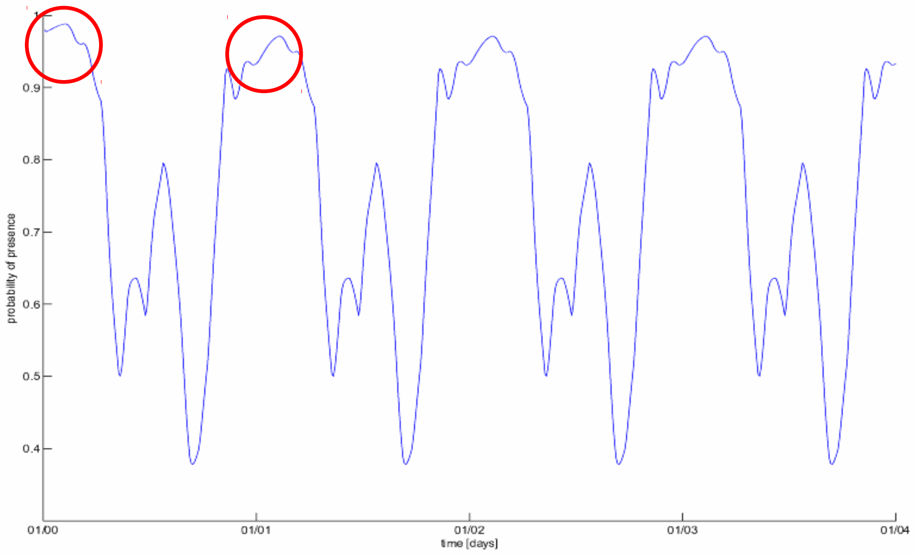


*Figure 8: Profile generation process*

*Fig. 9* shows a significant overall trend over this 30 peoples sample for the presence profiles at home. One can observe that individuals behave differently creating sometimes high variations. However, some kinks or breaks (emphasised in red circles) occur, they may result from the non-significant parameters expressed in the model (*cf. section 3.4.2*). Indeed, between two consecutive time-steps (e.g. 6:00am-7:00am & 7:00am-8:00am), the model changes (in the procedure, the utility parameters of the model are estimated at each time-step) with different significant and non-significant parameters and thus, the transition parameters of some individuals may substantially change between two adjacent time intervals, creating these kinks. However, the kinks do not exist for all individuals at these time values, indicating that this is only related to some parameters.

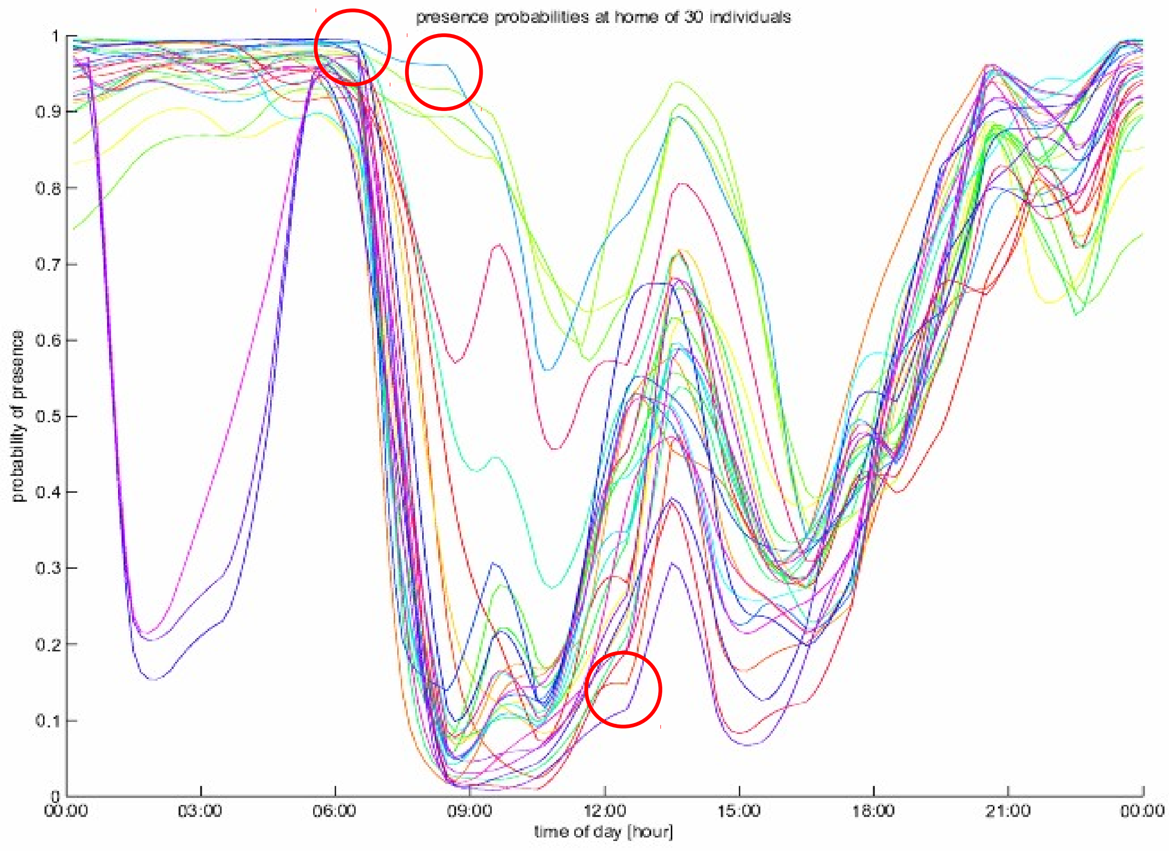


*Figure 9: 30 individual household presence profiles randomly chosen over the whole population*

## 4.2 Results

### 4.2.1 Household occupancy

The individual profiles of presence probabilities of the entire population are used to present here the statistical distribution of the resulting profiles. The following graphs show the results with the mean (black full line), the quantiles distribution (5% & 25% : dot & dash green lines, median in blue, 75% & 95% : dot & dash red lines) and the extrema values (min & max in thin dot line green and red respectively).

The first graph (*Fig.* 9) represents the resulting profiles for the overall set of individuals. An overall trend shows up and large variations are observed (e.g. difference between the presence profile for the 25% and 75% of the population is about 0.2).

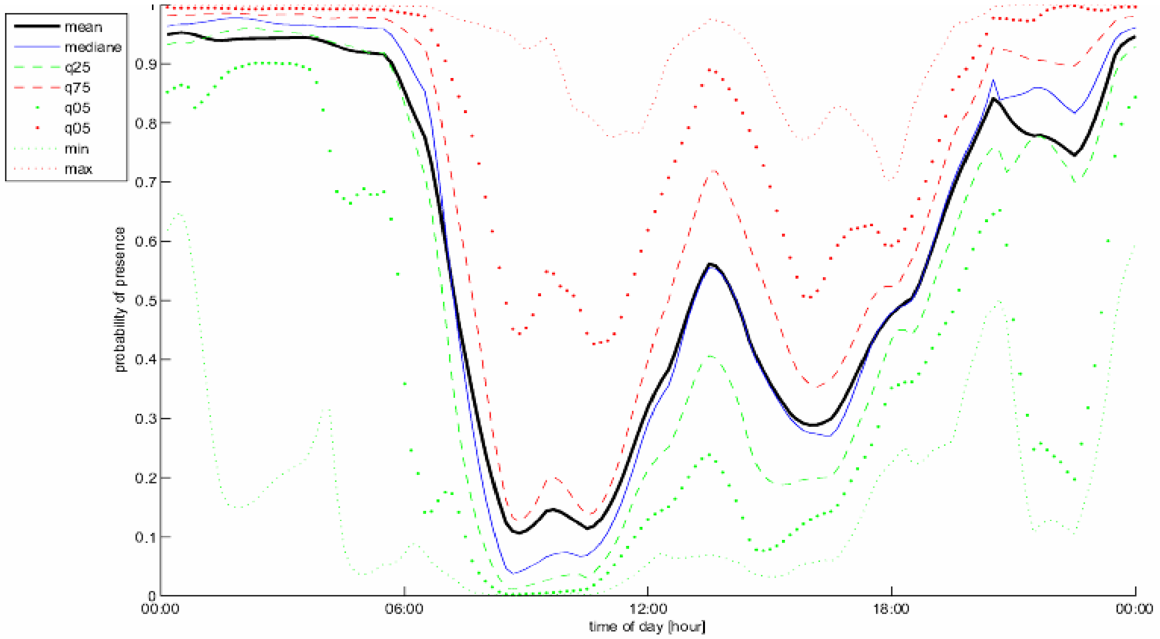


*Figure 10: presence probabilities profile for the whole population*

Then, a comparison is performed between two sub-populations concerning the presence during week days (here Monday to Friday). *Fig. 10* illustrates the presence profile of the active population (non-retired) for a week day, whereas f*ig. 11* shows the same profile for retired persons. It is globally observed that the retired people are a bit less at home around 9:00am but are significantly more present since 1:00pm to the end of the day.

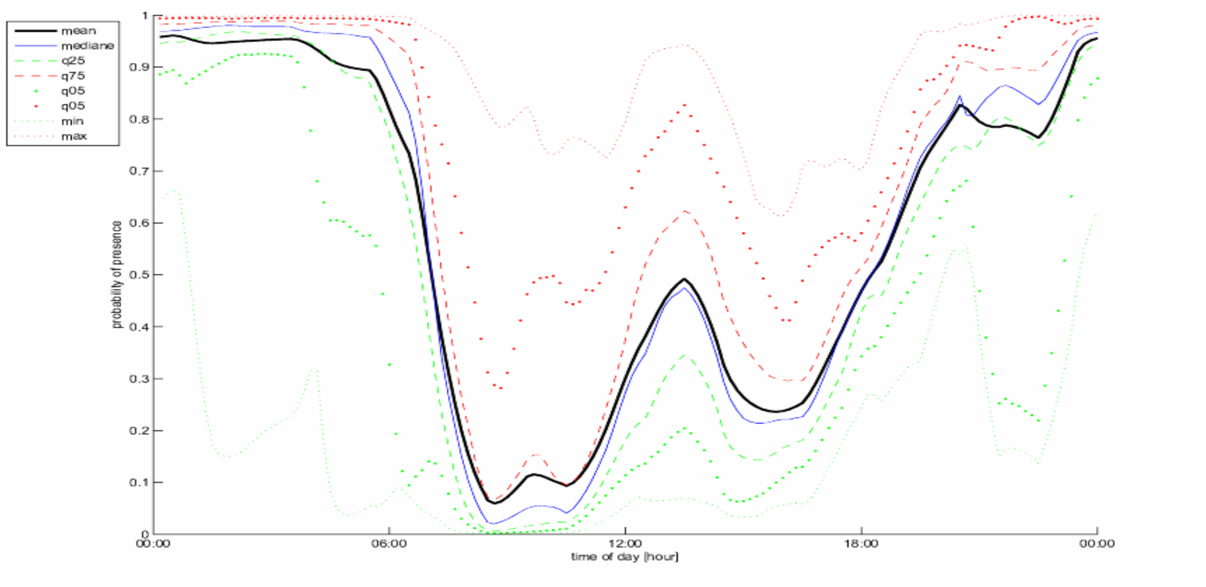


*Figure 11: presence probabilities profile for a week day for the active population*

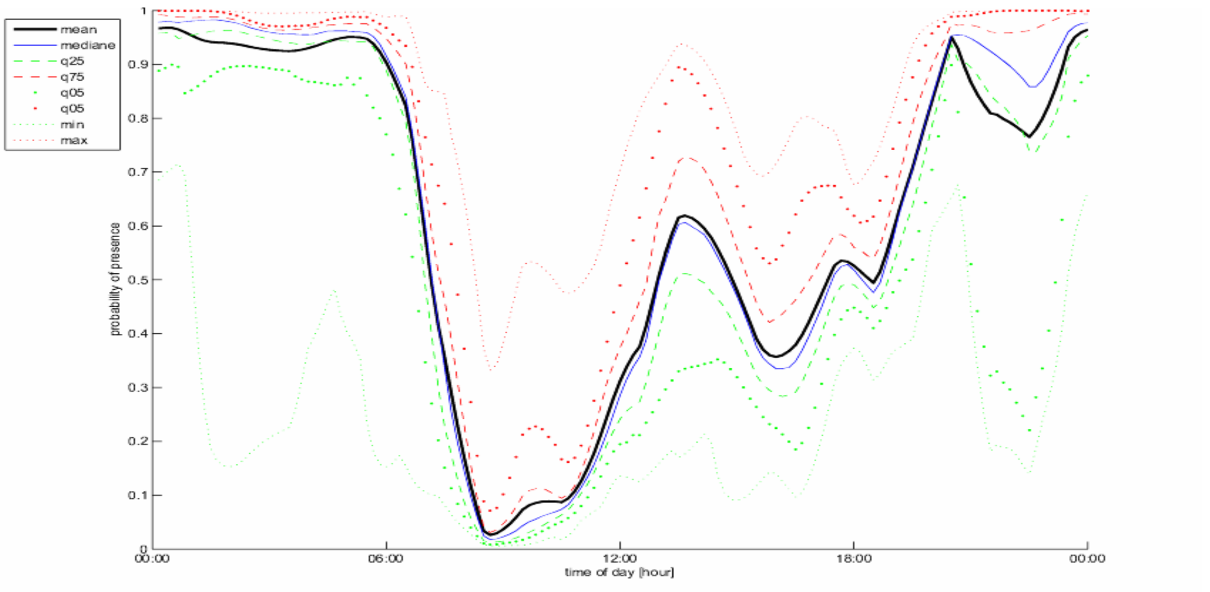


*Figure 12: presence probabilities profile for a week day for the retired population*

Next an another comparison is proposed between the household presence profiles of the men and the women (f*ig. 12 & fig.13* respectively). The two trends are similar, however one can notes that the men seem to be less at home than women especially during the night, since around 10:00pm until the morning around 7:00am, and also at lunch time.

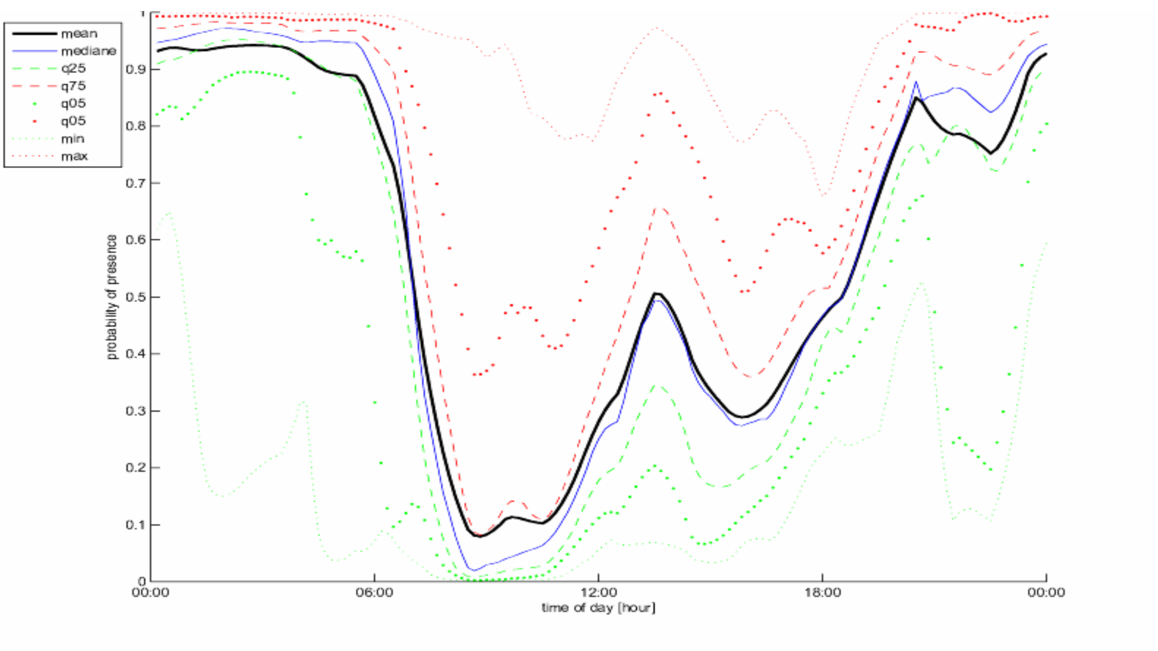


*Figure 13: presence probabilities profile for men*

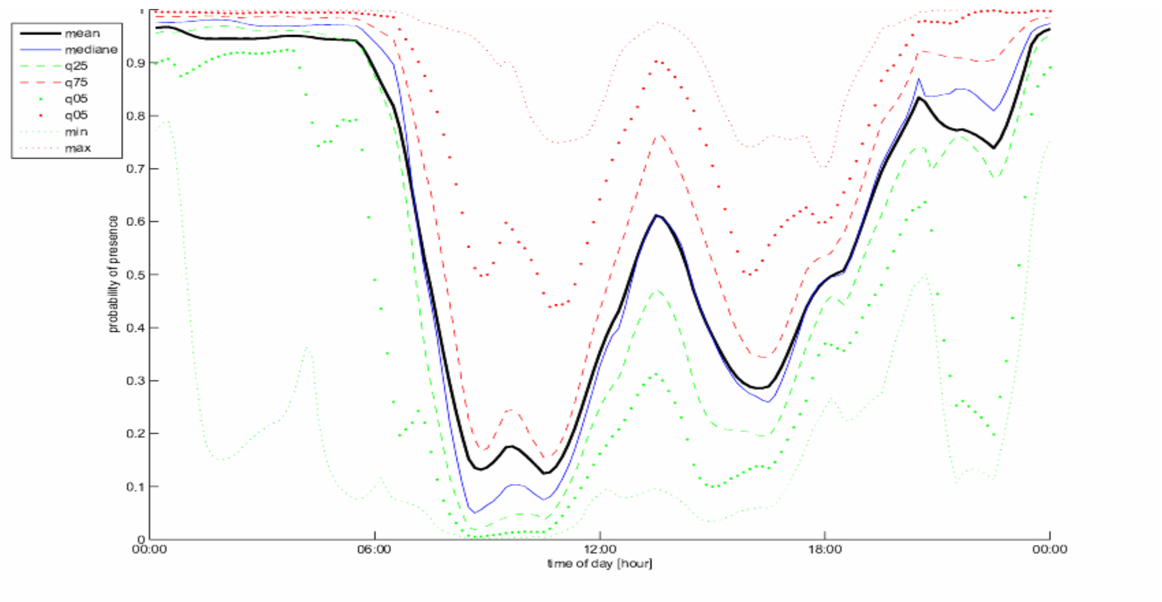


*Figure 14: presence probabilities profile for women*

### 4.2.2 Work-place occupancy

Similarly to the household occupancy results, the distributions of work-place occupancy results are presented here. First of all, the graph (f*ig. 15*) for the profiles of presence probabilities for the overall set of individuals is presented below. As it is attended, it is mostly complementary with the household occupancy case. The peak of presence at work-place is at 1:00pm which may appear surprising.

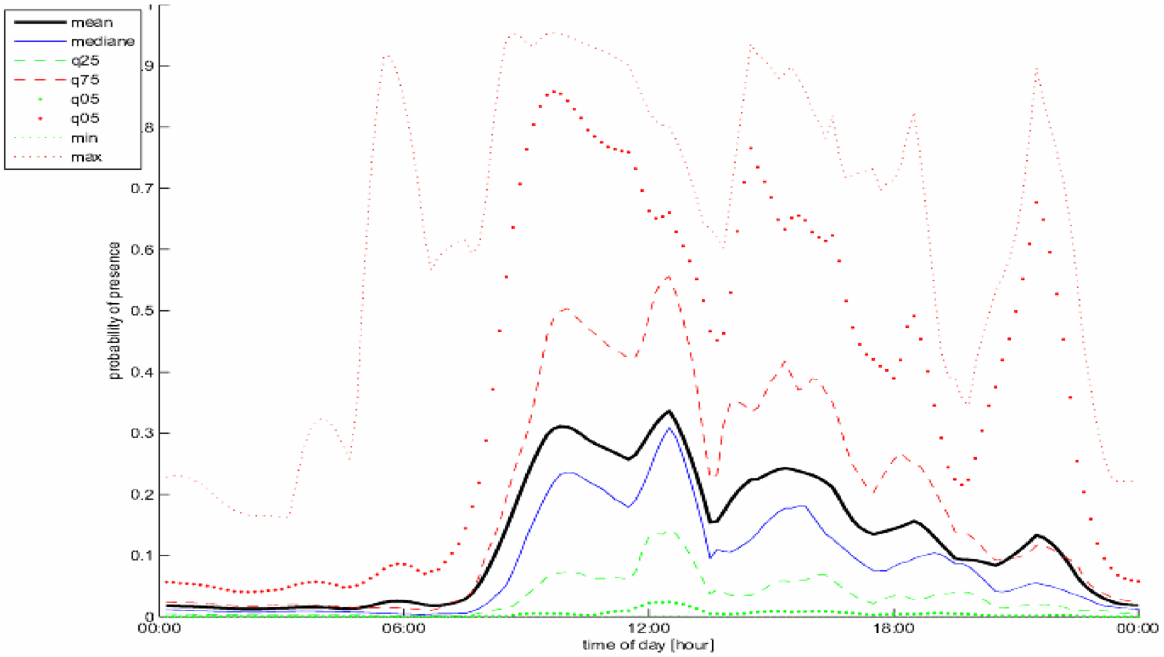


*Figure 15: presence probabilities profile for the whole population*

Next, the results for the presence on the work-place in a week day are presented below. The occupancy profile for the active population (*fig. 16*) follows the overall trend (f*ig. 15*) but with a higher magnitude (approx. +0.1 between 9:00am and 6:00pm). This seems consistent with expectations as it only considers working peoples during working days.

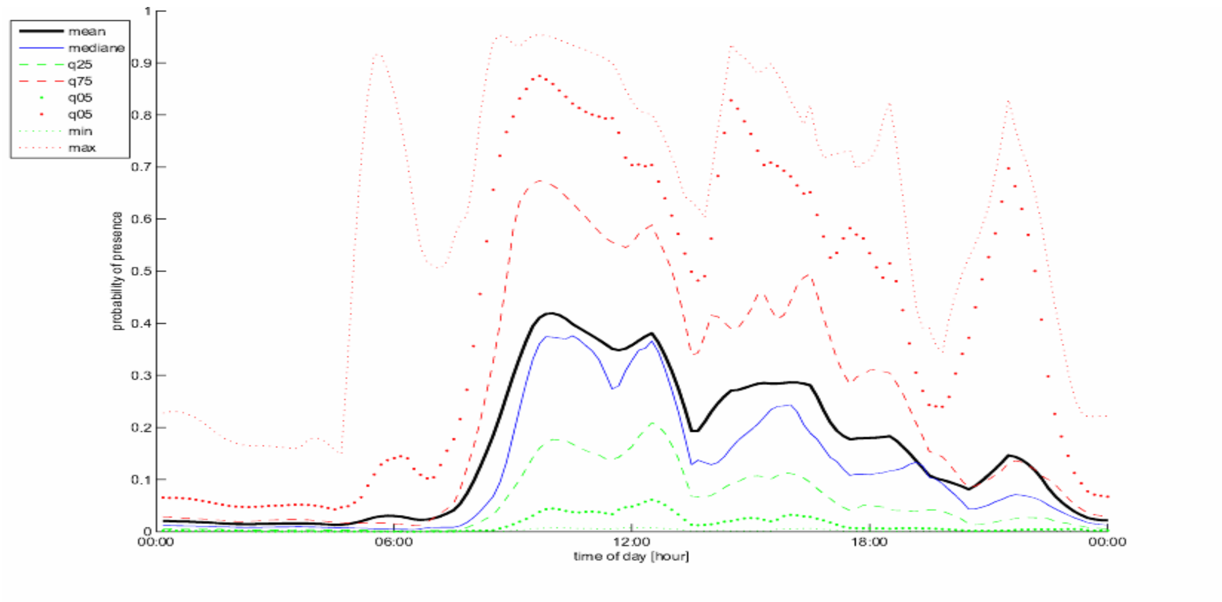


*Figure 16: presence probabilities profile for a week day for the active population*

However the results for the retired population (*fig. 17*) are not in line with the expectations with a large general magnitude and a peak of presence at work around 1:00pm. This is discussed later on (*cf. section 4.3*).

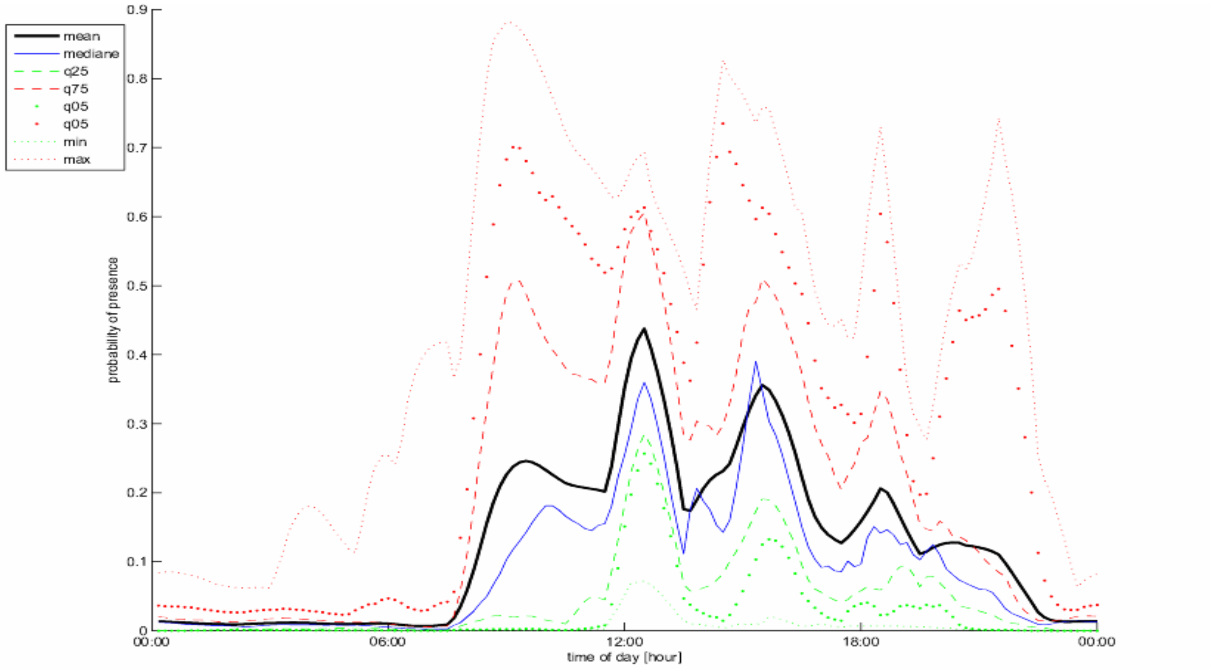


*Figure 17: presence probabilities profile for a week day for the retired population*

The following graphs show that the presence profiles at work-place for men and women (*fig. 18 & fig. 19* respectively) are largely different. These results show that the presence probabilities at work-place are almost twice as large for men than for women.

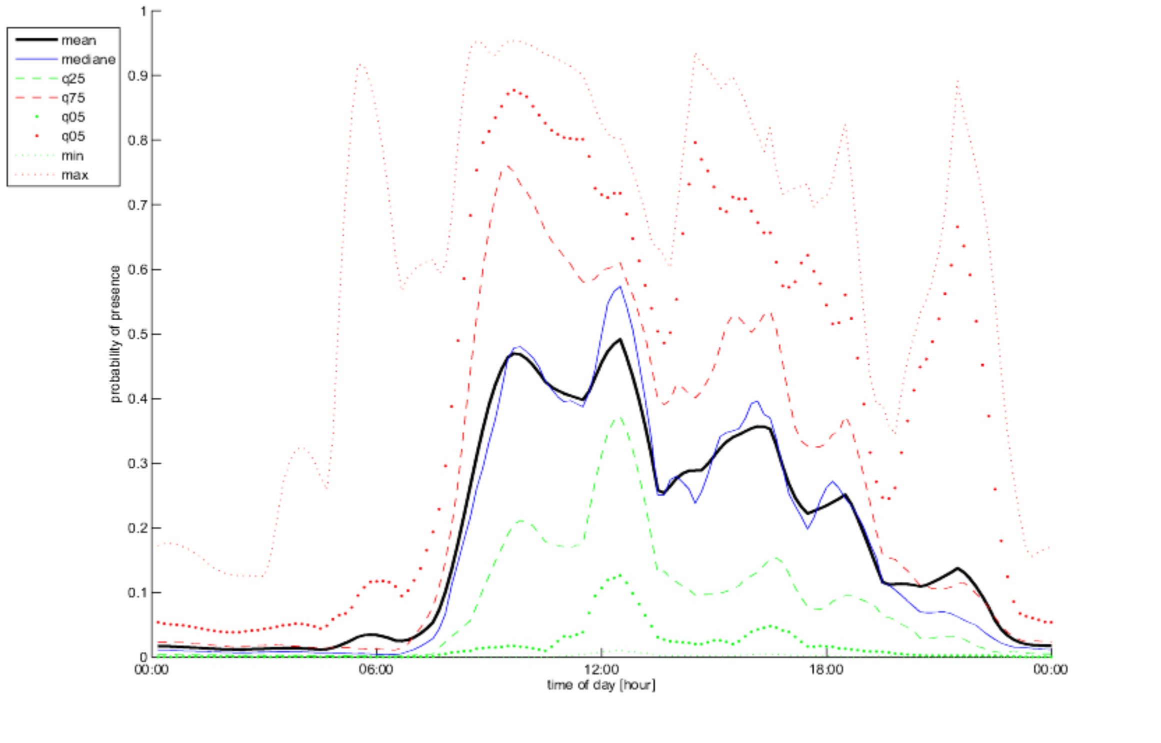


*Figure 18: presence probabilities profile for men*

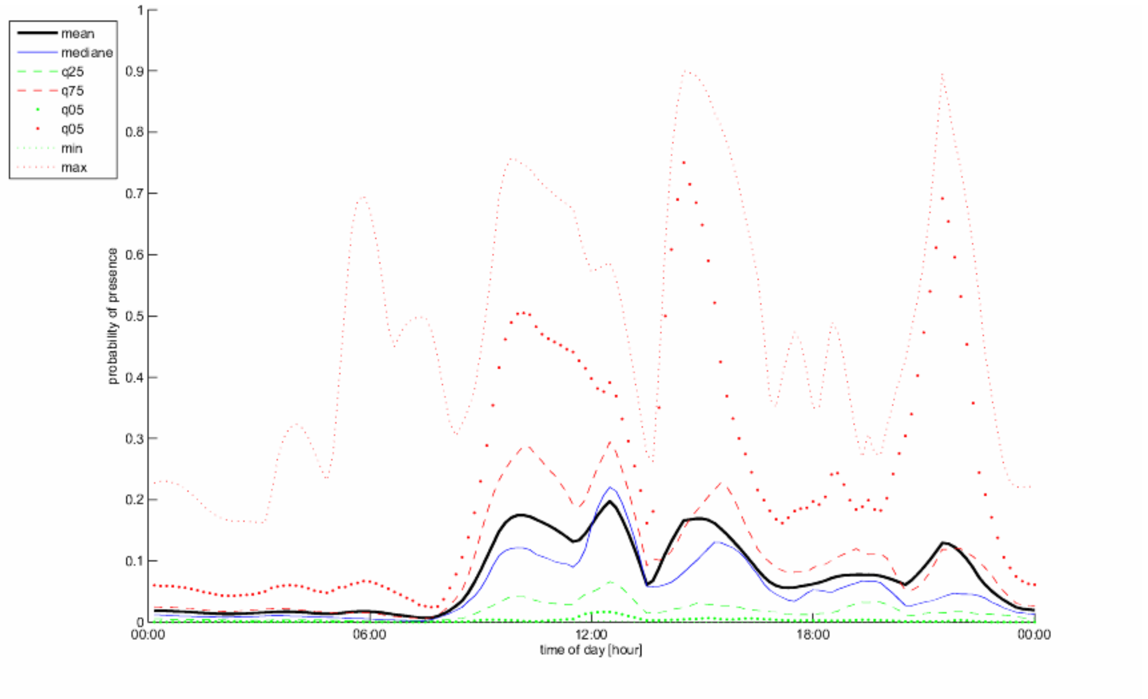


*Figure 19: presence probabilities profile for women*

## 4.3 Discussion

First of all, the results presented show general tendencies that are intuitive and seem to be close to reality, nevertheless, several kinks and breaks are observed in the graphs, especially in *fig. 18 showing* work-place occupancy of men. As discussed in *section 4.1*, this may be the consequence of the integration of non-significant utility parameters in the model, even if the presented results are showing the statistic distribution of all the individual profiles.

Concerning the work-place occupancy (*cf. section 4.2.2),* the overall (*fig. 15*) peak of presence around 1:00pm is probably due to the type of definition of the workplace in the TUS. There, also non-remunerated work is counted as a workplace. Thus, the peak is probably accentuated by this non-remunerated work, which is evident when looking at the course of the distribution of working retired persons (*fig. 17*), where the magnitude of this peak is even larger. In other words, the results of the presence probabilities during a week day of the retired population may appear contradictory but this is due to the interpretation of the TUS data [2]. Indeed in the latter, the type of location "work-place" does not necessarily mean a place for a paid work but may signify the place for doing voluntary work. That could explain the results presented in the working retired people presence profile (*fig. 17*).

## 4.4 Application

A major perspective for this study concerns its applications in building simulation. It was intended to apply the elaborated model, to identify the thermal influence of metabolic heat gains due to occupancy on a building (here a building of the EPFL) using  a dynamic simulation tool. Therefore, a specific profile of presence, supposed fitting with the EPFL working population, has been created. That corresponded to the work-place occupancy for active and higher-level educated persons (one for the week days and one for the weekend). The software CitySim was used to integrate the generated profile into the thermal simulation for the chosen building. The simulation input occupancy  information is given by an hourly profile of presence probalilities that is usually fed with standard values, e.g. values from the SIA[4] 2024 norm in the Swiss context.

Unfortunately, due to problems in the building geometry and limited available time, it was not possible to get correct results in time for this study.

## 4.5 Outlook

As discussed above (*cf. sections 4.1 & 4.3*), the main improvement fr the present project should be done in the implementation of the model regarding a better control of the non-significant utility parameters. Currently, because of a time constraint, manual checking and correction have been done concerning these non-significant parameters but this should be done systematically in order to enhance the quality of the model.

Additionnaly, the application case of the model for the thermal simulation of an EPFL building could be accomplished by correcting the geometry problems currently present.

[4] http://www.sia.ch

# 5 Conclusion

This project has started from scratch in the purpose of writing a model that predicts the occupancy as a function of individual characteristics and time of day from a bottom-up stochastic approach where no model was currently existing. The model has shown good results for predicting overall tendencies for the probabilities of presence. However several inaccuracies arose from non-significant utility parameters that have not been properly eliminated and from the TUS data that have presented some misinterpretations. So, in a future perspective, the next step should consist in establishing a systematic process that treat the last non-significant parameters, then the overall procedure may be simplified with the aim of being easily usable for many applications that are fully related to this project, especially in the energy optimisation of the building for heating, cooling and lighting needs.

# 6 References

[1] htpp://www.insee.fr

[2] K. Fisher et al., October 2011, Multinational Time Use Study - User’s Guide and Documentation

[3] Bierlaire, M., 2008. Estimation of discrete choice models with biogeme 1.8. Users manual.

[4] Fisher, K., Bennett, M., Tucker, J., Altintas, E., Jahandar, A., Jun, J.,other members of the Time Use Team, 2009. Technical Details of Time Use Studies. last updated 30 March 2009. Centre for Time Use Research, University of Oxford, United Kingdom.

[5] Bierlaire, M. (2003). BIOGEME: A free package for the estimation of discrete choice models, *Proceedings of the* 3rd Swiss Transportation Research Conference, Ascona, Switzerland.

# 7 Appendix

*Table 1: Corresponding signification of the parameter names used in the model*

| Parameter name (Biogeme) | corresponding dummy variable equal to one if condition holds (otherwise zero) |
|---|---|
| ASC1 | Alternative Specific Constant |
| | Age of the diarist |
| bage1 | <18 years |
| bage2 | 18 to 33 years |
| bage3 | 36 to 45 years |
| bage4 | 46 to 60 years |
| bage5 | 61 to 75 years |
| bage6 | > 75 years |
| | Age of the youngest child in the household |
| bagekidx-7 | no child |
| bagekidx1 | 0 to 4 years |
| bagekidx2 | 5 to 12 years |
| bagekidx3 | 13 to 17 years |
| bagekidx4 | 18 years and older |
| bcarer1 | caring of someone in the household |
| bcivstat2 | not in couple |
| bday0 | Week day (Monday to Thursday) |
| bday1 | Friday |
| bday2 | Weekend |
| | whether or not the diarist has a disability or long-term health limiting condition. |
| bdisab-8 | “unknown” |
| bdisab0 | no disability |
| bdisab1 | disability |
| | Education level of the diarist |
| bedtry1 | uncompleted |
| bedtry2 | secondary completed |
| bedtry3 | above secondary level |
| | Employment status |
| bempstat1 | full-time employed |
| bempstat2 | part-time employed |
| bempstat3 | unknown employment status |
| bempstat4 | not in paid work (retired, unemployed, ...) |
| | Family status |
| bfamstat0 | Adult aged 18 to 39 with no co-resident children <18 |
| bfamstat1 | Adult 18+ living with 1+ co-resident children aged <5 |
| bfamstat2 | Adult 18+ living with 1+ co-resident children 5-17, none <5 |
| bfamstat3 | Adult aged 40+ with no co-resident children <18 |
| bfamstat4 | Respondent aged <18 and living with parent(s)/guardian(s) |
| bfamstat5 | Respondent aged <18, living arrangement other or unknown |

| | |
|---|---|
| bhhtype1 | one person household-level |
| | Houshold type |
| bhhtype2 | couple household |
| bhhtype3 | couple + other persons household |
| bhhtype4 | other household types |
| | Monthly income level |
| bincorig1 | earning less than 700€ |
| bincorig2 | between 700€ and 1500€ |
| bincorig3 | between 1500€ and 3000€, |
| bincorig4 | more than 3000€, |
| bincorig99 | “doesn't know” |
| | Accommodation status |
| bownhome1 | own home |
| bownhome2 | rent home |
| bownhome3 | other |
| bretired1 | retired |
| bsex2 | female |
| bsingpar1 | single parent |
| bstudent0 | not a student |
| bstudent1 | student |
| burban2 | rural area |
| | Number of weekly working hours (with respect to the last representative week) |
| bworkhrs-1 | “not asked or no answer” |
| bworkhrs1 | 0 to 15 working hours |
| bworkhrs2 | 16 to 35 working hours |
| bworkhrs3 | 35to 45 working hours |
| bworkhrs4 | > 45 working hours |