Recognizing personalized flexible activity patterns

Author(s): Ordóñez Medina, Sergio A.

Publication Date: 2015-04

Permanent Link: https://doi.org/10.3929/ethz-b-000103800

Rights / License: In Copyright - Non-Commercial Use Permitted
Recognizing personalized flexible activity patterns

Sergio Arturo Ordóñez Medina

Future Cities Laboratory
Recognizing personalized flexible activity patterns

Sergio Arturo Ordóñez Medina  
Future Cities Laboratory (FCL)  
Singapore ETH Centre (SEC)  
138602 Singapore  
phone: +65-9298 7627  
fax: +41-44-633 10 57  
ordonez@ivt.baug.ethz.ch

April 2015

Abstract

In people’s daily plans, mandatory activities like rest, work or study, are already prearranged or fixed. Time planning or scheduling is therefore primarily to organize flexible activities and trips during the remaining time windows. Models of scheduling become powerful when they can be used to obtain macroscopic insights from microscopic changes. This paper presents a method to extract and model flexible activity patterns from real data, and to use them for activity chain generation. It proposes a multi-activity scheduling method which (i) doesn’t prioritize or fix any scheduling dimension, (ii) uses commonly available data (i.e. travel surveys and land use datasets) as its input, (iii) generates personalized solutions, and (iv) can be configured to be computed in tractable time for realistic time windows. Based on the concepts of known places and activity agenda, this method uses socio-demographic characteristics and matched behavioural parameters of the decision maker to schedule his/her flexible activities. To show the applicability and efficiency of this approach, flexible activity patterns from Singapore were extracted from a travel survey carried out in 2012. This survey contains reported motorized trips of 1% of the population during one day. A dataset with information of more than 100,000 activity locations, and estimated dynamic travel times from a large-scale agent-based transport model were employed for the calculations. The method was tested with a fraction of the survey observations which were not used for the estimation of the models. In order to assess the relevance of the personalized mechanisms of the model, the utility maximization algorithm was applied several times, comparing random locations versus systematically selected locations and randomly generated agendas versus systematically constructed agendas. Although in both cases optimal activity schedules are successfully generated, results show much better fit when personalized models are applied.
Keywords
Activity scheduling, Agenda, Discrete choice models, Utility maximization, Location choice

Preferred citation style
1 Introduction

In people’s daily plans, mandatory activities like rest, work or study, are already prearranged or fixed. Time planning or scheduling is therefore primarily needed to organize flexible activities and trips during the remaining time windows. Understanding how individuals from a certain city or region plan activities during these periods of time is very useful for many applications, such as travel demand modelling, land use planning and redevelopment, or market research and analysis. These models become a powerful tool when they can be used to perform thousands or millions of predictions leading to macroscopic insights from microscopic changes (e.g. changes in preferences or in the built environment). This paper presents a method to extract and model these flexible activity patterns from real data, and to use them to make activity chain predictions.

As mentioned by (Feil, 2010) the fundamental problem of activity scheduling is its combinatorial complexity due to its number of dimensions (activity durations, locations, number of activities, activity types, activity sequence, etc.). The modeller needs to develop solutions which allow solving the problem in tractable time. Complex methods and algorithms have been proposed to model these multi-activity scheduling decisions with spatio-temporal restrictions (Arentze and Timmermans, 2000, Doherty et al., 2002, Miller and Roorda, 2003, Arentze and Timmermans, 2004, Feil, 2010). Other methods are based on addressing only the next activity (Kuhnimhof and Gringmuth, 2009, Arentze and Timmermans, 2009) or performing continuous planning without taking into account activity locations (Märki et al., 2014). Many of these models are based on strong assumptions, such as a fixed number of the activities to be scheduled, or fixed activity durations, producing restricted results. One of the most common assumptions imposed in some heuristic or probabilistic models, such as Miller and Roorda, 2003, Arentze et al., 2010, Kuhnimhof and Gringmuth, 2009, is the prioritization of some scheduling dimensions during the decision process. That means, the scheduling process is carried out in a predefined and fixed order e.g. activity type -> location -> duration -> activity type -> etc. With this restriction, decisions in which a location is a priority cannot be modelled. This work proposes a multi-activity scheduling method which (i) doesn’t prioritize or fix any scheduling dimension, (ii) uses commonly available data (i.e. travel surveys and land use datasets) as its input, (iii) generates personalized solutions, and (iv) can be configured to be computed in tractable time for realistic time windows (although its complexity grows with the length of the time window). Each prediction is personalized, i.e. it uses socio-demographic characteristics and matched behavioural parameters of the decision maker. In the first step of the prediction, this information is used to identify an Activity agenda of flexible activities and a set of possible activity locations for them (choice set). An activity agenda is a collection of possible activities to be performed with a range of possible durations and frequencies as proposed by (Axhausen, 2006). In the second step a graph-based utility maximization algorithm takes the agenda, the collection of known locations, the locations where the decision maker plans to start and end his/her activities, and the available time budget, to find an optimal flexible activity chain (the prediction). This utility maximization algorithm is based on (i) the time geography
concepts introduced by (Hägerstrand, 1970) and developed by (Wilson, 2008), and (ii) the idea of using spatio-temporal graphs for activity-trip chain optimization proposed by (Arentze and Timmermans, 2004). This algorithm also uses the functions described by (Charypar and Nagel, 2005) to measure activity utilities and trip disutilities. Information about car-availability and dynamic travel times can be included in this algorithm, as they would affect the choices of location and transportation mode. If the method is told who the decision maker is, where he/she is, how much free time he/she has and where he/she has to be after that time, the model predicts a possible and detailed flexible activity chain which on average corresponds to the patterns of the city or region, and maximizes his/her utility. To show the applicability and efficiency of this approach, flexible activity patterns for Singapore were extracted from the Household interview travel survey carried out in 2012 (HITS 2012). This survey contains reported motorized trips of 1% of the population during one day. A dataset with information of more than one hundred thousand activity locations, and estimated dynamic travel times from a large-scale agent-based transport model implemented for Singapore (Erath et al. (2012)) were also used for more realistic calculations. To generate personalized activity agendas and location choice sets, Logistic regression models for each flexible activity type (eating, shopping, social activities, running errands, recreation), and for each place type (shops with high or low demand, eating places with high or low demand, community centres, home of others, parks, recreational spots) were estimated. These 13 models represent how socio-demographics and the locations where people live or work, are related to the flexible activities they perform (personalized patterns). Activity duration and travel time distributions were also extracted from the travel survey. The following section presents an overview of the whole method. The third section gives the details of the estimation of the Logistic regression models, while the fourth section elaborates on the activity agenda generation, the known places set formation, and the utility maximization algorithm. Section five describes differences between flexible activities observed and predicted, and predictions calculated using personalized agendas and locations are compared with optimal schedules generated with random inputs. Finally some conclusions and future work are discussed in the last section.
2 Overview

The overall objective of this work is to study how people plan their daily activities, to model these processes, and to use these models to schedule activities in different scenarios. In other words, the goal is to predict which sequence of activities will be planned by a certain person under some set of initial conditions. Accurate multi-activity scheduling is a non-trivial problem. First, these decisions depend on intrinsic variables related to the behaviour of the person who is planning, and on extrinsic variables related to the state of the rest of the world where the decision is taken. Second, multi-activity scheduling is a problem where many dimensions need to be addressed at the same time. Although a mere sequence of activities is sufficient for some applications, transport planning studies demand more dimensions. This work aims to schedule the sequence of the activities, their start times and durations, and the locations where the activities are performed. The mode of transportation of the trips carried out to perform activities at different locations is also predicted.

The concept of fixed and flexible activities has been used in several activity-based modelling studies (Doherty et al. (2002), Arentze and Timmermans (2000), Miller (2005), Chen and Kwan (2012)). These concepts allow studying activity scheduling at two levels. The first level is focused on the activity skeleton, which is composed of mandatory or fixed activities. The most important or primary human activities, like Home, Work and Study activities, belong to this level. Primary activities have been extensively studied by transportation modellers and urban planners (Hansen (1959), Small (1982), Vovsha and Gupta (2013), Ordóñez Medina and Erath (2013)). They try to find spatial patterns within geographical regions like cities, and temporal patterns during one day or longer periods of time, for example weeks. The second level is focused on non-mandatory or flexible activities. Most of the secondary human activities belong to this level. The challenge at this level is the variability of these activities, i.e. the number of combinations in which these activities are performed. This makes it more difficult to find a small number of patterns that can explain flexible activity decision making. For example, when analysing a travel survey containing daily activity chains of 1% of the population of a city, there are more than 900 combinations of activities. By contrast, there are just 40 combinations of primary activities after removing secondary activities. If primary activities in this dataset are just categorized as Home and Work (i.e. Studying is a type of work), only 10 combinations can be found. Thus, it seems a good strategy to study separately fixed and flexible activities to address this problem. This study is focused on flexible activities scheduling, assuming that fixed activities are already scheduled.

As mentioned in the previous section this multi-activity scheduling method is based on the concept of Activity agenda. An Activity agenda $A$ is a set of $n$ activities intended to be performed by a person. However, these activities are not set in time (start time and duration) or space (location). The information related to each activity $a_i$ is a typical
duration \(d\) and a typical frequency \(f\). These two attributes can be modelled as random variables, estimating probability distributions from observations. An Activity agenda provides a level of abstraction to model flexible activity decisions. It restricts the combinatorial problem from a universal set of possible activities to a controlled set of activities included in the agenda. Given an activity skeleton with fixed activities planned during a defined time period and time windows without planned activities, the agenda can be used to plan flexible activities during these free times. Models by (Nurul Habib and Miller, 2009) and (Nijland et al., 2012) also aim to generate personalized activity agendas with an econometric and a Bayesian approach respectively.

As this work also aims to resolve the locations where these flexible activities are performed, another concept must be introduced, that is the Set of known places. When a person plans flexible activities, it can be assumed that he/she only takes into consideration a limited number of activity locations. Known places is not the best name for this set, as a decision maker can know many more places which he doesn’t take into consideration or even is reluctant to travel to. Regarding the set of alternatives in a general choice, (Narayana and Markin, 1975) introduce 5 subsets: Unawareness set, Awareness set, Inept set, Inert set and Evoked set. Unawareness set and Awareness set are exclusive and, as the names imply, if the decision maker is aware of the alternative, it belongs to the awareness set, or otherwise to the Unawareness set. Inept set, Inert set and Evoked set are subsets of the Awareness set. If the decision maker has a negative image of an alternative it belongs to the Inept set, if the image is neutral to the Inert set and if it’s positive to the Evoked set. It is therefore necessary to point out that in this paper the term known places only refers to the locations where the decision maker is willing to travel to, i.e. locations that belong to the Evoked set in the decision maker mental map. Thus, a small set of \(m\) known places \(P\) is the second piece of input information given to this multi-activity scheduling method. For each known place \(p_i\), estimated travel times \(t_{ij}\) to all the other known places \(p_j\) in the set are assumed to be known. Each known place \(p_i\) also includes information of the flexible activities \(p_{aik}\) that can be performed there. Opening times could also be included, but won’t be mentioned in this paper for reasons of simplicity. In summary, the tuple \((A, P)\) is the model of the personalized information the decision maker uses when scheduling flexible activities. It is a limited representation of his/her mental map.

This work focuses on how activity agendas and known places can be extracted from common datasets. Figure 1 summarizes the processes developed (bright boxes), and the data used (dark boxes). The two dark boxes on the top represent the main datasets, with the spatial information (transportation network and activity facilities information) on the left whereas a travel or activity survey is on the right. If the estimations or measurements of multi-modal travel times between any pair of locations are not given, they can be calculated with a transportation network. If travel times by public transport are also needed, information about this system (stops, lines, transfers, etc.) must be provided. Information on activity facilities is also a key element. First, geographic locations are needed to calculate travel times. The type of the facility and the activities that can be performed there is essential when estimating sets of (known places). In general,
any information that reveals how attractive a place is to perform a certain activity is useful in this dataset (size, age, agglomeration level, etc.). The last main dataset is the travel or activity survey. It must contain the sequence of activities people from a certain geographical region perform during a defined period of time. Normally these datasets only incorporate one-day reports of a small sample of the population. Transportation modes of trips between locations where activities are taken place could be useful, but not necessary, because multi-modal travel times can be calculated using the other datasets.

The boxes with a thick border on the upper half of Figure 1: represent data and processes applied to a population dataset. Then, to extract temporal and spatial activity patterns of the population, these processes only need to be executed once. In contrast, the boxes at the bottom half represent the data and processes which must be applied for every activity scheduling, using the results from the population processes. As the main objective of this work is to explain how real data can be used for multi-activity scheduling, the five processes in the center of the figure will be described in detail. The "Go to place of type X" models are binary logistic models to predict how likely it is for a person to visit a certain type of place based on his/her socio-demographic characteristics. The "Perform
activity X” models are similar binary logistic models, but to predict how likely it is for a certain decision maker to perform a certain activity. These two sets of models, along with the extraction of activity duration and travel time distributions, generate the input parameters for the Selection of known places and the Activity agenda estimation. With this activity agenda, and the set of known places, optimal activity schedules can be found. The Spatio-temporal network method is a technique to find maximum-utility activity schedules. Although, with the activity agenda and the set of known places, realistic non-optimal schedules could also be identified using less expensive methods. In the next section, the binary logistic models are presented in detail. In addition, the Selection of known places, the Activity agenda estimation and the Spatio-temporal network method to calculate activity chains are elaborated in the third section. Finally, the results of applying the whole algorithm to a sample of people in the travel survey which were not used for the estimation of the models, are presented and analysed in the last section.

3 Participation in flexible activity purposes and awareness of place types

As mentioned above, this section presents in detail the development of the binary logistic models to extract activity and place type patterns from a population. These population models are represented by the "Go to place of type X" models and the "Perform activity X” models boxes included in the Figure 1: The main objective is to find out which socio-demographic and geographic attributes influence the type of places people from a certain geographical region travel to, and the activities they perform. These models are estimated with observations from a travel survey, a database of activity facilities, and estimated multi-modal travel times between these facilities. The following personal attributes from the Singaporean (Household interview travel survey) carried out in 2012 (HITS 2012) were included in the binary logistic models:

- Age
- Gender
- Car availability
- Ethnicity: The reported ethnicities are Chinese, Indian, Malay and others.
- Accessibilities: For each type of place, a measurement of accessibility was calculated from each primary activity locations of every person (i.e. residence, workplace, school, university, etc). The general accessibility of a person to a place type is the maximum accessibility among all his/her primary locations. Thus if for example, the workplace of a person is very accessible to shopping places, but the home residence is not, the shop-accessibility measurement for that person is still high. Each measurement was calculated with the following formula: \[ \text{Acc}(p_i, t) = \sum_{p_j \in P(t)} (A(p_j)^{\beta}) e^{\alpha T(p_i, p_j)} \]. In this equation \( t \) is a type of place, \( P(t) \) is the set of facilities of type \( t \), \( T(p_i, p_j) \) is the travel time between the
facilities $p_i$ and $p_j$, and $A(p_j)$ is the attraction level of the $j$th facility. Finally, $\alpha$ and $\beta$ are parameters calibrated to optimize the linear relation of the accessibility measurement with the duration of activities performed at each place type.

- Size of household: Number of people in the household
- Role in the household: Four roles were defined. A main role is assigned to the person with highest income in the household. A partner role is assigned to the person with most similar age to the person with the main role. Members younger and older than the main and partner persons are assigned to the younger and older roles respectively.
- Income: Net income of the person
- Main income: The income of the person with highest income within the household
- Home time: The time the person stays at home during the day
- Work time: The time the person spends working or studying during the day

The place types included in the method are recreational facilities, parks, community centres, homes of others, high-demand shopping places, low-demand shopping places, high-demand eating places, and low-demand eating places. The categorization of a place as high demand or low demand is determined by its size and/or the number of trips reaching that place according to the travel survey. One place (location or building) can be classified into multiple types. Flexible activities included are eating, shopping, social meetings, recreational activities, and running errands. A binary logistic model was developed for each type of place, as well as for each flexible activity. With 8 place types and 5 activities, a total of 13 models were estimated. Each trip to a flexible activity in the travel survey can represent an observation (a set of input variables and an outcome) for these models. For each trip the respondent reports the activity purpose and the place type of his/her destination. In the case of the "Go to place of type $X$" models, the outcomes are based on the relation between activities and place types shown in Figure 2.

This relation was defined according to the number of trips to perform a specific activity at a specific place type in the travel survey. For example there are 4 observations in the survey where people go to a Recreational facility to run errands. But compared with almost 200 observations of this activity at shopping malls, that number is negligible. To categorize binary observations of the "Go to place of type $X$" models, the outcomes are based on the relation between activities and place types shown in Figure 2. This means that for a certain place type, the observations from the survey are filtered by the activity type, and just activity types related to this place type are taken into account for the logistic regression. The table shows some examples of the classification of trips

1. When one of the types of the reported trip destination is $X$, the outcome is 1.
2. When the purpose of the reported trip (the activity) is related to the type $X$, but the destination of the trip is not of type $X$, the outcome is 0.
3. When the destination of the trip is not of type $X$, but the reported purpose of the trip (the activity) is not related to the type $X$, the observation is not counted.

This means that for a certain place type, the observations from the survey are filtered by the activity type, and just activity types related to this place type are taken into account for the logistic regression. The table shows some examples of the classification of trips
Figure 2: Relation between flexible activities and place types

<table>
<thead>
<tr>
<th>Place type</th>
<th>Purpose</th>
<th>Age</th>
<th>Gender</th>
<th>Outcome</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park</td>
<td>rec</td>
<td>40</td>
<td>Female</td>
<td>0</td>
<td>Rec is related to community centre</td>
</tr>
<tr>
<td>Community centre</td>
<td>work</td>
<td>35</td>
<td>Male</td>
<td>N/A</td>
<td>Work is not a flexible activity</td>
</tr>
<tr>
<td>Shopping place</td>
<td>shop</td>
<td>45</td>
<td>Male</td>
<td>N/A</td>
<td>Shop is not related with community centre</td>
</tr>
<tr>
<td>Community centre</td>
<td>social</td>
<td>35</td>
<td>Male</td>
<td>1</td>
<td>Social is related with community centre</td>
</tr>
</tbody>
</table>

Table 1: Some examples of the classification of trips for the "Go to community centre" model.

for the "Go to community centre" model.

On the other hand, the binary classification for the "Perform activity X" models are much simpler. If a person travels to perform a certain flexible activity during the reported period (e.g. one day), that person is counted as a positive observation for the corresponding activity model. If the person doesn’t travel to perform the current flexible activity during the reported period, the outcome of the observation is 0. The "Perform activity X" binary
Choice models are the main input for the agenda estimation.

The objective at this point is to implement 13 binary classifiers that will return a binary output $y$ (two classes) given a vector of inputs $X$. For this experiment 90% of the data was used for estimation or training and 10% was for testing. Binary logistic regression needs significant manual work from the modeller, specifying utility functions for the alternatives. Although extra initial work is necessary, the results of the estimation determine which input variables are more significant or less significant to predict the output. It also gives a probability for each alternative when a new vector of inputs $X$ is evaluated. The significance levels of the input variables are important to better understand the accuracy and draw better conclusions for decision makers. In the next subsections these models are summarized and analysed.

### 3.1 "Go to place of type X" model results

As mentioned before the place types modelled were recreational facilities, parks, community centres, homes of others, high-demand shopping places, low demand shopping places, high-demand eating places, and low-demand eating places. Figure 3 shows some of the dependencies of socio-demographic characteristics on travelling to or not to certain place types. Using these figures, 8 linear utility functions were formulated. They represent the influence of some of the mentioned personal attributes on travelling or not to each place type. The utility of going to high-demand shopping places is presented in the equation 1:

$$U_{shop} = \beta_{ageL17(ageL17)} + \beta_{ageH17(ageH17)} + \beta_{fem(fem)} + \beta_{carAv(carAv)} + \beta_{chi(chi)} + \beta_{ind(ind)} + \beta_{mly(mly)} + \beta_{accLow(accLow)} + \beta_{accHigh(accHigh)} + \beta_{hSize(hSize)} + \beta_{mainR(mainR)} + \beta_{prtnR(prtnR)} + \beta_{yngR(yngR)} + \beta_{inc0(inc0)} + \beta_{inc(inc)} + \beta_{hIncL7k(hIncL7k)} + \beta_{hIncM7k(hIncM7k)} + \beta_{timeH(timeH)} + \beta_{timeL7h(timeL7h)} + \beta_{timeM9h(timeM9h)} + \beta_{noWork(noWork)} + \epsilon$$

The variables of this function are listed in Table 2.

Logistic regression was used to estimate the coefficients. Table 2 presents the estimated coefficients of the "Go to high-demand shopping places" model with their significance levels. Similar utility functions were estimated for the other 7 place types. For the sake of brevity table 3 summarizes the significance of the 11 socio-demographic attributes in the 8 "Go to place of type X" models. For the coefficient estimation, dummy variables were created for categorical variables while linear scaling was applied to continuous
Figure 3: Some dependencies on the willingness to travel or not to travel to places of different types, based on socio-demographic attributes. The height of each bar represents the fraction of the number of observations which the person goes to the corresponding place type, over the total number of observations in the corresponding category. The color of each bar represents the absolute number of observations in the corresponding category, the more transparent the color the more observations.

Variables to limit their ranges to the interval [0, 1].

Interesting findings can be seen in this table. Age influences almost all the place types, with a strong negative correlation with recreation facilities, parks and home of others; and a strong positive correlation with community centres. Accessibility is one of the strongest predictors, with positive correlations with almost all the place types, except for low-demand eating places. Although this negative correlation was not expected at first glance, one could explain this arguing that people might prefer to go to inaccessible restaurants because of privacy or that low-demand restaurants tend to be located at less accessible areas. Time at home is also a strong predictor with strong correlation with parks, community centres and home of others. Finally, it’s interesting to find out that the household size has a negative correlation with travelling to home of others, which possibly means that households or families with a bigger number of members carry out more activities at their own living places.
Table 2: Coefficient estimates of the "Go to high-demand shopping places" model. Shaded rows mark significant variables.

3.2 "Perform activity X" models estimation

Flexible activity types in this study include eating, shopping, social meetings, recreational activities, and running errands. Similar to the "Go to place of type X" models, Figure 4 illustrates some of the dependencies of socio-demographic characteristics on whether or not performing certain flexible activities during the period of time of the survey. Linear utility functions were also employed for each of the 5 models to represent these dependencies. Equation 2 models the utility of the "Perform eating" model, Table 4 presents its estimated coefficients, and Table 5 summarises the priorities of the 11 attributes in the performance or non-performance of these 5 flexible activities. For the coefficient estimation, dummy variables were created for categorical variables and linear scaling was applied for continuous variables to fix their ranges to the interval [0, 1].
Table 3: Summary of the "Go to place of type X" choice models absolute impact. In this table, green represents a positive dependency while red refers to a negative one. Blue color means that the relation between the parameter and the outcome is not always increasing or decreasing. Deeper shades represent higher impact or higher absolute value of the utility parameter. The letter "s" means that the variable is significant according to the t-test. Yellow cells are significant categorical variables. The text in these cells represents the order from the most to the least significant category. For ethnicity "c" means Chinese, "i" refers to Indian, "m" is for Malay and "o" means other. For household role "m" is for the main role, "p" means the partner role, "o" stands for the older role and "y" is the younger role.

$$U_{eat} = \beta_{ageL40}(ageL40) + \beta_{ageH40}(ageH40) + \beta_{fem}(fem) + \beta_{carA}(carA) + \beta_{chi}(chi) + \beta_{ind}(ind) + \beta_{mly}(mly) + \beta_{accEatHigh}(accEatHig) + \beta_{accEatLow}(accEatLow) + \beta_{accShopHigh}(accShopHigh) + \beta_{accShowLow}(accShowLow) + \beta_{accHomeO}(accHomeO) + \beta_{hSize}(hSize) + \beta_{mainR}(mainR) + \beta_{prtnR}(prtnR) + \beta_{yngR}(yngR) + \beta_{inc0}(inc0) + \beta_{inc}(inc) + \beta_{hIncL2k}(hIncL2k) + \beta_{hIncM2k}(hIncM2k) + \beta_{timeH}(timeH) + \beta_{timeW}(timeW) + \beta_{noWork}(noWork) + \epsilon$$

The variables of this function are listed in Table 4. Similar to "Go to place of type X" models, interesting findings can be drawn from the results summarized in Table 5. Age is a very strong predictor with positive correlation with all the flexible activities. As expected working time is a strong negative predictor, affecting 4 of the 5 flexible activities. For each activity the accessibilities to places of related types (according to Figure 2) were included as possible predictors. It’s interesting to notice that the accessibility to high-demand shopping places and community centres
Figure 4: Some dependencies on the performance or non-performance of flexible activities based on socio-demographic attributes. The height of each bar represents the fraction of the number of observations which the person performs the corresponding activity, over the total number of observations in the corresponding category. The color of each bar represents the absolute number of observations in the corresponding category, the more transparent the color the more observations.

is highly and negatively correlated with performing errands. Having a car resulted in negative correlation with shopping. Finally, an analysis of the dependencies of the household roles in shopping and social activities was made. The results showed that the older role performs more these activities than all the others roles, while the younger role performs the least activities; and, although the main role performs more shopping activities than his/her partner, the partner role performs more social activities.
<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
<th>Std err</th>
<th>t-test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>-3.84</td>
<td>0.3</td>
<td>-12.8</td>
<td>0.00</td>
</tr>
<tr>
<td>Age below 40</td>
<td>3.04</td>
<td>0.437</td>
<td>6.95</td>
<td>0.00</td>
</tr>
<tr>
<td>Age above 40</td>
<td>1.92</td>
<td>0.252</td>
<td>7.62</td>
<td>0.00</td>
</tr>
<tr>
<td>Female</td>
<td>-0.087</td>
<td>0.0572</td>
<td>-1.51</td>
<td>0.13</td>
</tr>
<tr>
<td>Car availability</td>
<td>0.284</td>
<td>0.0598</td>
<td>4.75</td>
<td>0.00</td>
</tr>
<tr>
<td>Chinese ethnicity</td>
<td>0.113</td>
<td>0.138</td>
<td>0.81</td>
<td>0.42</td>
</tr>
<tr>
<td>Indian ethnicity</td>
<td>-0.0933</td>
<td>0.154</td>
<td>-0.6</td>
<td>0.55</td>
</tr>
<tr>
<td>Malay ethnicity</td>
<td>-0.138</td>
<td>0.157</td>
<td>-0.88</td>
<td>0.38</td>
</tr>
<tr>
<td>Other ethnicity</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Acc. eat high</td>
<td>3.47</td>
<td>0.182</td>
<td>19.08</td>
<td>0.00</td>
</tr>
<tr>
<td>Acc. eat low</td>
<td>1</td>
<td>0.3</td>
<td>3.33</td>
<td>0.00</td>
</tr>
<tr>
<td>Acc. shop high</td>
<td>1.74</td>
<td>0.129</td>
<td>13.42</td>
<td>0.00</td>
</tr>
<tr>
<td>Acc. shop low</td>
<td>-1.68</td>
<td>0.32</td>
<td>-5.27</td>
<td>0.00</td>
</tr>
<tr>
<td>Acc. home other</td>
<td>0.553</td>
<td>0.171</td>
<td>3.24</td>
<td>0.00</td>
</tr>
<tr>
<td>Household size</td>
<td>1.04</td>
<td>0.19</td>
<td>5.5</td>
<td>0.00</td>
</tr>
<tr>
<td>Main role</td>
<td>0.275</td>
<td>0.105</td>
<td>2.61</td>
<td>0.01</td>
</tr>
<tr>
<td>Partner role</td>
<td>0.0994</td>
<td>0.0874</td>
<td>1.14</td>
<td>0.26</td>
</tr>
<tr>
<td>Younger role</td>
<td>0.0592</td>
<td>0.125</td>
<td>0.48</td>
<td>0.63</td>
</tr>
<tr>
<td>Older role</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>No income</td>
<td>0.0464</td>
<td>0.0963</td>
<td>0.48</td>
<td>0.63</td>
</tr>
<tr>
<td>Personal income</td>
<td>-0.256</td>
<td>0.264</td>
<td>-0.97</td>
<td>0.33</td>
</tr>
<tr>
<td>Household income</td>
<td>-1.79</td>
<td>0.474</td>
<td>-3.79</td>
<td>0.00</td>
</tr>
<tr>
<td>Household income below 2.5k</td>
<td>-0.0902</td>
<td>0.215</td>
<td>-0.42</td>
<td>0.67</td>
</tr>
<tr>
<td>Household income above 2.5k</td>
<td>-2.29</td>
<td>0.192</td>
<td>-11.95</td>
<td>0.00</td>
</tr>
<tr>
<td>Time at home</td>
<td>-2.29</td>
<td>0.192</td>
<td>-11.95</td>
<td>0.00</td>
</tr>
<tr>
<td>Time at work</td>
<td>-4.5</td>
<td>0.614</td>
<td>-7.33</td>
<td>0.00</td>
</tr>
<tr>
<td>No working time</td>
<td>-0.743</td>
<td>0.154</td>
<td>-4.83</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 4: Coefficient estimates of the "Perform eating" model. Shaded rows mark significant variables.
Table 5: Summary of the "Perform activity X" choice models absolute impact. In this table, green represents a positive dependency while red refers to a negative one. Blue color means that the relation between the parameter and the outcome is not always increasing or decreasing. Deeper shades represent higher impact or higher absolute value of the utility parameter. The letter "s" means that the variable is significant according to the t-test. Yellow cells are significant categorical variables. The text in these cells represents the order from the most to the least significant category. For ethnicity "c" means Chinese, "i" refers to Indian, "m" is for Malay and "o" means other. For household role "m" is for the main role, "p" means the partner role, "o" stands for the older role and "y" is the younger role.
4 Known places, activity agenda and optimal scheduling

In this section, three of the processes mentioned in Figure 1 of Section 2 will be explained in detail. The *Distance-based selection of known places* generates a controllable number of destinations for a given decision maker. The *Activity agenda estimation* yields a set of flexible activities which are not yet defined in time or space, as explained in the section 2. These two processes have to be executed once for each decision maker. Finally, on the bottom of figure 1, the *Spatio-temporal network method to calculate activity chains* returns an optimal sequence of flexible activities to be performed in a given time window. This process has to be executed for every multi-activity scheduling. One person can have multiple scheduling episodes during the defined period of time (e.g. one day).

4.1 Known places set construction

This procedure returns a fixed number $N$ of known places for a decision maker, based on his/her socio-demographic attributes which were listed in the beginning of section 3. It also returns travel times between each pair of known places. The *selection of known places* depends on the person and his/her socio-demographic characteristics and not in the specific situation of this person when he/she is making the trip decision. Previous models like Timmermans et al. (1982), Zheng and Guo (2008) or Horni (2013) estimate specific choice sets depending on the current situation of the traveller (e.g. the location of the origin). Their personalized choice sets can also be used once.

This paper proposes two steps for the *selection of known places*. The first selection depends only on the types of the places, selecting $F \times N$ places ($F$ is a factor greater than 1). The second step depends on the geographical dimension and the selection returns $N$ out of $F \times N$ places. Next the two steps are explained in detail.

4.1.1 Step 1: Selection by type

With the utility functions estimated with the "Go to place of type" models, described in the previous section, the probability of knowing places of any type can be estimated. It is important to point out again that the definition of "knowing" a place is that the decision maker is willing to travel there. Also, each location can be classified into multiple types, making it more attractive. The probability that a person $y$ knows places of type $t$ is given
by the equations (3):

\[ U_t = \beta_{t0}(v_{y0}) + \beta_{t1}(v_{y1}) + \ldots + \beta_{tm_t}(v_{ym_t}) + \epsilon_t \]

\[ y_t = \begin{cases} 1, & \text{if } U_t > 0 \\ 0, & \text{if } U_t \leq 0 \end{cases} \]

\( \epsilon_t \sim \text{Logistic} \)

\[ P(y_t = 1) = \frac{1}{1 + e^{-(\beta_{t0}(v_{y0}) + \beta_{t1}(v_{y1}) + \ldots + \beta_{tm_t}(v_{ym_t}))}} \]

Where \( U_t \) is the utility of knowing a place of type \( t \), \( \beta_{ti} \) is the \( i^{th} \) coefficient estimated for the type \( t \), \( v_{yi} \) is the \( i^{th} \) characteristic of the person \( y \), \( m_t \) is the number of related variables, \( \epsilon_t \) is the unknown utility which is assumed to have a logistic distribution, \( y_t \) is the binary random variable which represents person \( y \) knows or does not know places of type \( t \), and \( P(y_t = 1) \) is the probability of this person knows places of this type.

As a destination can be classified into multiple types, the probability for a person to know a destination is given by the equation (4):

\[ P(d_y = 1) = 1 - \prod_{t \in T(d)} (1 - P(y_t = 1)) \]

In this equation \( T(d) \) is the set of types of the destination \( d \), \( d_y \) is the binary random variable which represents the person \( y \) knows the destination \( d \), and \( P(d_y = 1) \) is the probability of this person to know this destination. The product term represents the joint probability of not knowing the location by any of its types. Thus, the only way of not knowing a place is because the person doesn't know it by any of its types, and a place with many types is more probable to be known.

This probability represents how much a person is attracted to travel to a certain destination according to its types. Thus, the Selection by type calculates these probabilities for all the destinations included in the region of interest, and randomly selects \( F \times N \) destinations using these values.

4.1.2 Step 2: Selection by travel time to primary locations

It is assumed that short travel times for a person between his/her primary locations and a destination raises the probability that a person "knows" that destination. In other words,
Distribution of the travel time to destinations ratio

Figure 5: Distribution of the travel time to destinations ratio. For all observations of a person \( y \) travelling to a destination \( d \), this graph shows the histogram of the ratio between the minimum travel time from any primary location of \( y \) to \( d \), and the travel time from \( y \)’s residence to \( y \)’s work or study location. People tend to travel to locations near their primary activity locations. Figure 5 shows the distribution of the ratio of the minimum travel time from any primary location of a person \( y \) to a destination \( d \), to the travel time from \( y \)’s residence to \( y \)’s work or study location. When this ratio is less than 0.5, it means that travel time from \( d \) to one of the primary locations is less than half of that from home to work. A great majority of destinations are in this range.

Based on this assumption, this second step is about selecting destinations according to the travel times from the primary locations of the person to the destination. Thus, a destination \( d \) of type \( t \) is more likely to be known by a person \( y \) if the travel time from the primary locations \( P(y) \) to \( d \) fits better with the general travel time distribution of the type \( t \), as explained below. The extracted travel time density plots for each place type are presented in the figure 6.

Given a certain travel time \( tt_d \) from a primary location to a destination \( d \) of type \( t \), a measurement \( f_t(tt_d) \) can be obtained after evaluating the corresponding kernel density estimate. This number represents how much \( tt_d \) fits the type \( t \). As a person \( y \) can have multiple primary locations, multiple \( f_t \) can be calculated to \( d \). The maximum of these \( f_t \) values represents the geographic correspondence of \( d \) for \( y \). The maximum means that one of \( y \)’s primary locations is the reason to know \( d \). If \( d \) is classified in more than one...
type, the maximum of multiple $f_{t1}, f_{t2}, ...$ will be the final measurement.

Finally, a random selection of $N$ destinations can be made using their corresponding $f$ values.

### 4.2 Activity agenda selection

As mentioned in previous sections an agenda $A$ is a set of $n$ activities intended to be performed by a person $y$. These activities are not yet scheduled (start time and duration) or located. The information related to each activity $a_i$ is a duration $d$ and a frequency $f$. The probability that a person $y$ performs a flexible activity $a$ is given by the equation 5:

$$U_a = \beta_{a0}(v_{a0}) + \beta_{a1}(v_{a1}) + ... + \beta_{ama}(v_{ama}) + \epsilon_a$$

$$y_a = \begin{cases} 1, & \text{if } U_a > 0 \\ 0, & \text{if } U_a \leq 0 \end{cases}$$

$$\epsilon_a \sim \text{Logistic}$$

$$P(y_a = 1) = \frac{1}{1 + e^{-(\beta_{a0}(v_{a0})+\beta_{a1}(v_{a1})+...+\beta_{ama}(v_{ama}))}}$$
In this equation $U_a$ is the utility of performing the activity $a$, $\beta_{ai}$ and $\nu_{ai}$ are the $i^{th}$ coefficient and variable related to the activity $a$, $m_a$ is the number of related variables, $\epsilon_a$ is the unknown utility which is assumed to have a logistic distribution, $y_a$ is the binary random variable which indicates performing or not performing the flexible activity $a$, and $P(y_a = 1)$ is the probability of performing this activity.

Using the utility functions estimated in the section 3, the probability of the performance of any flexible activity $a$ by person $y$ can be calculated. As just respondents who reported flexible activities were used to validate the method, the probability of performing each activity is conditional. The definition of conditional probability and the corresponding correction is presented in the equation 6.

$$P(y_a = 1|y_f = 1) = \frac{P(y_a = 1 \cap y_f = 1)}{P(y_f = 1)} = \frac{P(y_a = 1)}{P(y_f = 1)}$$

(6)

Where $y_f$ is the binary random variable which represents the performance or non-performance of any flexible activity, and $P(y_f = 1)$ is the probability of performing any flexible activity.

After calculating these probabilities for a person $y$ a random selection was performed for each flexible activity to construct the agenda. Finally, the duration and frequency distributions of each flexible activity were extracted from the travel survey and assigned to the corresponding activities in each agenda.

4.3 Spatio-temporal network method to calculate activity chains

The final step of this algorithm is the generation of a fully specified activity chain. The developed algorithm finds an approximately optimal activity chain according to its utility of performing activities and its dis-utility of travelling, under some restrictions. The inputs include the origin $o$ and destination $d$, start time $st$ and latest end time $et$, the Activity agenda and the Set of known places. From the origin node $(o, st)$, a spatio-temporal network is recursively constructed inside the corresponding space-time prism. Each node is then defined by a geographic location and a time stamp. Nodes are only created at known places and at specific times according to a time bin. This controls the size of the network. Trips and activities are represented by links of the network. To assign costs to the activity links an extension of the activity utility function presented in Charypar and Nagel (2005) was employed. This version, presented in Equation 7 and illustrated in Figure 7, incurs a penalty if same activity is performed more than once during a period of time. Hence, the utility of a link within a path depends on the previous links from the origin. This restricts the information sharing between paths with common
Activity performing utility as a function of the its duration and the time lapsed since the last time the same activity was performed.

Travel disutility functions from (Charypar and Nagel, 2005) were also employed for the travel link costs, using the travel times between known places.

Let $U(t_d, t_l)$ be the utility of performing an activity with duration $t_d$ and time lapsed since the last time the activity was performed $t_l$.

$$U(t_d, t_l) = \beta T_x \left( \frac{t_d + g(t_l)}{T_0 + g(t_l)} \right)$$

The penalty function $g(t_l)$ is defined as:

$$g(t_l) = \begin{cases} \alpha e^{\gamma(T-t_l)} - \alpha, & \text{with } t_l < T \\ 0, & \text{with } t_l \geq T \end{cases}$$

$$\gamma = \frac{1}{T} \ln \left( \frac{T + \alpha}{\alpha} \right)$$

Thus:

$$g(t_l)t_l = \begin{cases} 0, & \text{when } t_l = T \\ T, & \text{when } t_l = 0 \end{cases}$$

Where $t_d$ is the duration of the activity, $T_x$ is the typical duration, $T_0$ is the minimal duration, $\beta$ is the marginal utility of performing the activity, $t_l$ is the time lapsed since the last time that activity was performed, $g(t_l)$ is its penalty function, $\alpha$ is a parameter to determine how strong is the applied penalty (see Figure 7), and $\gamma$ is just a function of $\alpha$.

The shortest path from the origin node $(o, st)$ to the destination node $(d, et)$ can be obtained as a result of the network construction by saving the costs of the paths while the network is recursively growing. This path models an optimal and fully characterized activity chain of flexible activities, including the number of the activities, the order, the start times, the durations and the places where the activities are performed. Transportation modes of the trips between the locations are also defined. Figure 8 illustrates an example.
Figure 8: Shortest path of activities and trips found by the multi-activity scheduler.
5 Results and analysis

The main objective of the experiments carried out running the described method, is to demonstrate the importance of personal characteristics when scheduling flexible activities. Hence, the following results will show that predictions made using socio-demographics and geographical information are more accurate than random methods. As mentioned before, 10% of the travel survey respondents were taken aside from the beginning of the process for validation purposes. A total of 566 persons were scheduled using the algorithm described in this paper. These predictions took 31 seconds using a one-thread java program running on a windows machine with a common processor (intel i7) and running one thread; the full population of Singapore would take 8.6 hours with this rate. Furthermore, for each person in the test set, other maximum utility activity schedules were calculated, but using different selection methods for the known places and randomly constructed agendas. Figure 9 presents the result of comparing these methods with the real schedules. For the chart at the top, flexible activity durations were extracted from the reported trips, from random input predictions, and from the developed systematic predictions. Systematic predictions were more accurate, with very short activity durations for the random draws. Although the same activity utility function was used for both estimates, with the same duration distribution for each flexible activity, the number of activities predicted with random inputs was higher. These results were expected because the agendas were not restricted for the random input predictions and the utility grows when many short activities are scheduled (due to the performing activity utility function). On the other hand, restricted agendas and the extension of the performing activity utility function, which penalizes continuous activities of the same type, generate shorter activity chains with longer activity durations. The second bar chart shows that the number of predicted activities were quite similar, except for running errands with 40% less observed activities.

Figure 10 presents a crucial issue in transportation studies, travel times. This figure shows comparisons of travel time distributions when going to eat on the top, and when going to shop at the bottom. These comparisons are directly related to the selection of known places. Four selection methods were implemented to define 20 known places for each decision maker. The first method selects the known places randomly (i), the second selects places according to travel time distributions (ii), the third is the 2 step method described before (iii), and the fourth selects the 20 best locations according to type and travel time distributions (iv)(deterministic). Full random selection (i) can not reproduce the observed distributions at all (as expected), while optimal deterministic selections (iv) fails reproducing the long tail of the observed distribution. In contrast using travel time distributions (ii) and (iii) from the training data (5100 people) reproduces the observed distributions of the 566 testing records better. The results also suggest that the selection by place type (first step of the proposed selection) doesn’t improve the prediction significantly.
Figure 9: Comparison of maximum utility predictions made with socio-demographics and geographical information against maximum utility predictions using random inputs.
Travel time distributions to eat using different known place selection methods

![Travel time distributions to eat using different known place selection methods]

Travel time distributions to shop using different known place selection methods

![Travel time distributions to shop using different known place selection methods]

Figure 10: Travel time kernel densities of eating and shopping activities. Prediction capabilities after using four different known place selection methods can be compared.
Figure 11: Prediction comparison of the total number of recreational activities and activity performance share by age. For each interval the share is obtained dividing the number of people with a recreational activity scheduled over the total number of people in that interval.

Finally to assess the correlation of socio-demographic characteristics on the predicted schedules, Figure 11 compares the random input prediction with the systematic prediction at 10 different age intervals. The systematic method is very accurate judging by the high correlation of age in performing recreational activities and going to community centres. On the other hand, the random input prediction is just related with the number of respondents at each age interval. This can be appreciated comparing the total number of activities on the left with the activity performance share on the right.
6 Conclusions and outlook

In this paper, personalized flexible activity scheduling was studied. One of the main goals was to use commonly available data. Flexible activity patterns of the city of Singapore were extracted, estimating 13 *Binary logistic regression* models from the *Household interview travel survey* carried out in 2012. Results successfully show that using socio-demographic and geographical characteristics of people, as an input of a maximum-utility activity scheduler, improves prediction capabilities.

The *Activity agenda* concept was employed to restrict the activity scheduling problem. It was demonstrated that with systematic agendas a more accurate number of activities was predicted than using random constructed agendas. More accurate activity durations were also achieved, as the restriction on the type of flexible activities resulted in longer and fewer activities scheduled.

After comparing four selection methods of flexible activity destinations (choice set), neither random selection, or optimal selection were capable to represent observed travel time distributions. The best approach was achieved by the two step selection method described in this paper; first a selection by type according to individual characteristics of the decision maker, and second a random selection according to observed travel time distributions.

Comparing with other activity scheduling methods, this approach doesn't have enough prediction capabilities yet. More research must be carried out on the activity utility function as it depends in just one number, the typical duration. Current successful binary classifiers, such as *AdaBoost*, *Support Vector Machines* or *Random Forest*, can be tested to extract flexible activity patterns with non-linear correlations.

On the computation issue, a duration of 8 hours predicting flexible activities of a full population is a tractable time. Furthermore, the method is fully parallelizable as the proposed activity scheduling for one person doesn’t depend on other people. If shorter computation times are needed the activity scheduling algorithm can be relaxed, assuming that the perform activity utility doesn’t depend on previous activities. In this case the Dijkstra algorithm can be applied to find the optimal activity-trip chain, and the complexity would drop from $O(n^2)$ to $O(n \log(n))$. 
7 Acknowledgements

The author would like to thank the following Singaporean Authorities for providing access to data and valuable review: Land Transport Authority, Urban Redevelopment Authority, Singapore Land Authority (SLA Digitised Land Information). Financial support comes from the National Research Foundation (NRF) of Singapore and ETH Zurich research fund.
References


