Simulating Work-Leisure Cycles in Large Scale Scenarios
Models and Implementation

Author(s):
Ordóñez Medina, Sergio A.

Publication Date:
2016

Permanent Link:
https://doi.org/10.3929/ethz-a-010833127

Rights / License:
In Copyright - Non-Commercial Use Permitted
SIMULATING WORK-LEISURE CYCLES IN LARGE SCALE SCENARIOS: MODELS AND IMPLEMENTATION

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

SERGIO ARTURO ORDÓÑEZ MEDINA

Master of Science, Universidad De Los Andes, Bogotá, Colombia

born on 08.01.1986

citizen of

Colombia

accepted on the recommendation of

Prof. Dr. Kay W. Axhausen, examiner
Prof. Dr. Marcela Munizaga, co-examiner
Dr. Alex Erath, co-examiner

2016
1 Introduction
  1.1 Motivation .................................................. 1
  1.2 Research goal ............................................. 2
  1.3 Structure of the document ................................. 3
  1.4 Declaration of contributions .............................. 4

2 Overview
  2.1 Activity-based transport models and agent-based mobility
       simulations ................................................. 5
  2.2 MATSim framework .......................................... 7
  2.3 Multi-Activity scheduling problem ......................... 8
  2.4 Singapore scenario ......................................... 10
  2.5 Proposed multi-day simulation ............................ 12

3 Multi-agent transport simulation for a multi-day time horizon 15
  3.1 Standard MATSim multi-day simulation ..................... 15
    3.1.1 Multi-day transport demand in MATSim ............... 15
    3.1.2 Multi-day transport supply in MATSim ............... 16
    3.1.3 MATSim modules in a multi-day simulation .......... 17
  3.2 MATSim extension overview ................................ 19
    3.2.1 Fixed and flexible activities ......................... 19
    3.2.2 Activity scheduling on-the-fly ....................... 20
  3.3 Passive planning ........................................... 21
    3.3.1 Overview ................................................ 21
## Contents

3.3.2 Passive planning model ........................................ 22
3.4 Application of Passive Planning .................................. 26
3.4.1 Passive Planning with travel time information .............. 26
3.4.2 Passive Planning: Multi-day scenario with an unexpected event ........................................... 27
3.5 Conclusion ........................................................... 31

4 Multi activity scheduler with a time budget ......................... 33
4.1 Overview .............................................................. 34
4.2 Spatio-temporal graph based scheduling algorithm ............ 35
4.2.1 Input information .................................................. 35
4.2.2 Spatio-temporal graph definition ............................... 36
4.2.3 Graph construction and optimal path detection algorithm .................................................. 39
4.3 Initial tests with controlled inputs .................................. 44
4.3.1 Marginal utility for activity performing study case ........ 44
4.3.2 Effect of transportation mode costs on activity scheduling study case ........................................ 44
4.4 Using the algorithm within an agent-based simulation ....... 47
4.5 Summary and conclusions ........................................... 52

5 Primary activity weekly patterns of public transport users in Singapore .................................................. 55
5.1 Type of activity detection ............................................. 58
5.1.1 Household interview travel survey summary .................. 58
5.1.2 Discrete choice models for activity type ....................... 59
5.2 Weekly primary activity patterns .................................... 69
5.2.1 Extracting activities from public transport smart card data in Singapore ........................................ 69
5.2.2 Inference of the main weekly primary activity patterns in Singapore ............................................ 71
5.3 Conclusions and discussion .......................................... 77

6 Personalized flexible activity scheduling using Singapore data ...................................................... 79
6.1 Overview of the flexible activities scheduling method ......... 81
6.2 Participation in flexible activity purposes and awareness of place types ............................................. 83
6.2.1 "Go to place of type X" model estimation ...................... 89
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>MATSim evolutionary algorithm.</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>Summary of the MATSim Singapore scenario development.</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>MATSim population model extension for passive planning.</td>
<td>22</td>
</tr>
<tr>
<td>3.2</td>
<td>New agent types for passive planning.</td>
<td>23</td>
</tr>
<tr>
<td>3.3</td>
<td>Planning engine and Planner manager relationship.</td>
<td>25</td>
</tr>
<tr>
<td>3.4</td>
<td>Planning flexible activities using updated travel times vs. euclidean distance.</td>
<td>27</td>
</tr>
<tr>
<td>3.5</td>
<td>Big shops in the Singapore scenario.</td>
<td>28</td>
</tr>
<tr>
<td>3.6</td>
<td>Number of trips to shop and eat by day of the week.</td>
<td>30</td>
</tr>
<tr>
<td>3.7</td>
<td>Number of trips to shop at different location types by day of the week.</td>
<td>30</td>
</tr>
<tr>
<td>3.8</td>
<td>Travel time distributions of trips to shop originated in the affected area by day of the week.</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>Proposed activity performing utility function.</td>
<td>39</td>
</tr>
<tr>
<td>4.2</td>
<td>Case study 1: Sensitivity of the optimal activity-travel path to the marginal utility of performing home activity.</td>
<td>45</td>
</tr>
<tr>
<td>4.3</td>
<td>Case study 2: Activity-trip chain scheduled by the proposed algorithm.</td>
<td>47</td>
</tr>
<tr>
<td>4.4</td>
<td>Computation time per iteration of MATSim evolutionary algorithm.</td>
<td>50</td>
</tr>
<tr>
<td>4.5</td>
<td>Evolution of the daily utilities of agents within the MATSim evolutionary algorithm.</td>
<td>51</td>
</tr>
<tr>
<td>5.1</td>
<td>Method to identify weekly activity patterns of public transport users in Singapore.</td>
<td>57</td>
</tr>
<tr>
<td>5.2</td>
<td>Logarithmic rate of the number of observations of students’ activities by duration.</td>
<td>63</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>5.3</td>
<td>Logarithmic rate of the number of observations of students’ activities by start time.</td>
<td>64</td>
</tr>
<tr>
<td>5.4</td>
<td>Logarithmic rate of the number of observations of workers’ activities by duration.</td>
<td>65</td>
</tr>
<tr>
<td>5.5</td>
<td>Logarithmic rate of the number of observations of workers’ activities by start time.</td>
<td>66</td>
</tr>
<tr>
<td>5.6</td>
<td>Main weekly work activity patterns for public transport users in Singapore.</td>
<td>75</td>
</tr>
<tr>
<td>5.7</td>
<td>Main weekly study activity patterns for public transport users in Singapore.</td>
<td>76</td>
</tr>
<tr>
<td>6.1</td>
<td>Processes for multi-activity scheduling using Singapore datasets.</td>
<td>82</td>
</tr>
<tr>
<td>6.2</td>
<td>Descriptive analysis of socio-demographic attributes in HITS.</td>
<td>86</td>
</tr>
<tr>
<td>6.3</td>
<td>Relation between flexible activities and place types.</td>
<td>88</td>
</tr>
<tr>
<td>6.4</td>
<td>Dependencies of place types on socio-demographic attributes.</td>
<td>90</td>
</tr>
<tr>
<td>6.5</td>
<td>Summary of the &quot;Go to place of type X&quot; models estimation.</td>
<td>92</td>
</tr>
<tr>
<td>6.6</td>
<td>Dependencies of flexible activities on socio-demographic attributes.</td>
<td>94</td>
</tr>
<tr>
<td>6.7</td>
<td>Summary of the &quot;Perform activity X&quot; models estimation.</td>
<td>95</td>
</tr>
<tr>
<td>6.8</td>
<td>Distribution of the travel time to destinations ratio from HITS 2012.</td>
<td>98</td>
</tr>
<tr>
<td>6.9</td>
<td>Travel time kernel density plots of each place type extracted from HITS 2012.</td>
<td>99</td>
</tr>
<tr>
<td>6.10</td>
<td>Activity duration distribution by type from HITS 2012.</td>
<td>100</td>
</tr>
<tr>
<td>6.11</td>
<td>Example of selected evoked places of a decision maker according to his/her primary locations.</td>
<td>102</td>
</tr>
<tr>
<td>6.12</td>
<td>Comparison of reported vs. predicted activity duration and number of activities.</td>
<td>103</td>
</tr>
<tr>
<td>6.13</td>
<td>Predicted travel time distributions for the five flexible activities.</td>
<td>104</td>
</tr>
<tr>
<td>6.14</td>
<td>Prediction comparison of the total number of recreational activities and activity performance share by age.</td>
<td>105</td>
</tr>
<tr>
<td>6.15</td>
<td>Zipf probability mass function of activity chains in HITS and predicted ones.</td>
<td>106</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Weekly activity demand generation processes for the Singapore scenario.</td>
<td>112</td>
</tr>
<tr>
<td>7.2</td>
<td>Real vs. estimated distribution of workers in the testing data.</td>
<td>116</td>
</tr>
<tr>
<td>7.3</td>
<td>Real vs. estimated distribution of students in the testing data.</td>
<td>117</td>
</tr>
<tr>
<td>7.4</td>
<td>HITS vs. estimated start time distribution of the home activity by day of the week.</td>
<td>119</td>
</tr>
<tr>
<td>7.5</td>
<td>HITS vs. estimated duration distribution of the home activity by day of the week.</td>
<td>119</td>
</tr>
<tr>
<td>7.6</td>
<td>HITS vs. estimated start time distribution of the work activity by day of the week.</td>
<td>120</td>
</tr>
<tr>
<td>7.7</td>
<td>HITS vs. estimated duration distribution of the work activity by day of the week.</td>
<td>120</td>
</tr>
<tr>
<td>7.8</td>
<td>HITS vs. estimated start time distribution of the study activity by day of the week.</td>
<td>121</td>
</tr>
<tr>
<td>7.9</td>
<td>HITS vs. estimated duration distribution of the study activity by day of the week.</td>
<td>122</td>
</tr>
<tr>
<td>7.10</td>
<td>Relation between flexible activities and place types in Singapore.</td>
<td>124</td>
</tr>
<tr>
<td>7.11</td>
<td>Travel time distribution comparison between observed car trips in HITS and selected Evoked places.</td>
<td>130</td>
</tr>
<tr>
<td>7.12</td>
<td>Travel time distribution comparison between observed public transport trips in HITS and selected Evoked places.</td>
<td>131</td>
</tr>
<tr>
<td>7.13</td>
<td>Histogram of P.T. departures in the simulation vs. CEPAS before calibration.</td>
<td>134</td>
</tr>
<tr>
<td>7.14</td>
<td>Computation time of the iterations of the weekly simulation.</td>
<td>135</td>
</tr>
<tr>
<td>7.15</td>
<td>Evolution of agents’ utility during 100 iterations.</td>
<td>136</td>
</tr>
<tr>
<td>7.16</td>
<td>Evolution of the car histogram from iteration 0 to 20.</td>
<td>137</td>
</tr>
<tr>
<td>7.17</td>
<td>Share of flexible activities by day of the week: HITS vs. Simulation.</td>
<td>138</td>
</tr>
<tr>
<td>7.18</td>
<td>Spatial comparison of shopping in Singapore.</td>
<td>139</td>
</tr>
<tr>
<td>7.19</td>
<td>Spatial comparison of medical activities in Singapore.</td>
<td>140</td>
</tr>
<tr>
<td>7.20</td>
<td>Start time distributions of eating activities: HITS vs. Simulation.</td>
<td>142</td>
</tr>
<tr>
<td>7.21</td>
<td>Duration distributions of shopping activities: HITS vs. Simulation.</td>
<td>143</td>
</tr>
<tr>
<td>7.22</td>
<td>Histogram of departures by public transport after calibration: CEPAS vs. Simulation.</td>
<td>144</td>
</tr>
</tbody>
</table>
List of Figures

9.1 Start time distributions of shopping: HITS vs. Simulation. 161
9.2 Start time distributions of running errands: HITS vs. Simulation. 162
9.3 Start time distributions of recreational activities: HITS vs. Simulation. 162
9.4 Start time distributions of religious activities: HITS vs. Simulation. 163
9.5 Start time distributions of medical activities: HITS vs. Simulation. 163
9.6 Duration distributions of eating: HITS vs. Simulation. 164
9.7 Duration distributions of running errands: HITS vs. Simulation. 165
9.8 Duration distributions of recreational activities: HITS vs. Simulation. 165
9.9 Duration distributions of religious activities: HITS vs. Simulation. 166
9.10 Duration distributions of medical activities: HITS vs. Simulation. 166
9.11 Spatial comparison of running errands in Singapore. 167
9.12 Spatial comparison of eating in Singapore. 168
9.13 Spatial comparison of recreation in Singapore. 169
9.14 Spatial comparison of religion in Singapore. 170
List of Tables

1.1 Contributions of the author in previously published papers: 4

3.1 Marginal utilities of flexible activities used in the test scenario: 28

4.1 Case study 3: Comparison conditions between Standard MATSim and Passive Planning: 49

5.1 Number of observations in HITS 2012 by day of the week: 59

5.2 Types of smart cards used in Singapore and the way the types are recorded: 60

5.3 Resulting parameters of the study model utility functions: 67

5.4 Resulting parameters of the work model utility functions: 67

5.5 Resulting extra parameters to estimate impact of the week day on the activity type: 69

5.6 Number of identified public transport consistent activities by type and day of the week: 71

6.1 Calibrated parameters for accessibility calculation: 85

6.2 Some examples of the classification of trips for the “Go to community centre” model: 89

7.1 Classification of locations from collected points of interest: 123

7.2 Tested accessibility definitions: 126

7.3 Performance of activity type random forest models: 127

7.4 Performance of activity type random forest models: 129

7.5 Real vs. Predicted matrix of the recreational locations model: 129

7.6 Initial conditions of MATSim weekly simulation: 132

7.7 Calibrated mode parameters of weekly simulation: 133

7.8 Flexible activity frequencies for the weekly simulation: 135
Abstract

Multi-agent transport simulation models are tools to support decision making. Evaluations and comparisons of several future scenarios can be performed using this approach. With the development of information technologies and computational power, it is now possible to simulate the transport system of an urban population in order of magnitude of $10^7$ agents. MATSim is a large-scale multi-agent transport simulation model based on human activities. It can simulate in parallel the mobility of millions of agents with reasonable computational effort. However, as the period of MATSim simulations is only one day, multi-day mobility patterns cannot be studied. This restricts the analysis of various transportation planning issues. For instance, recent studies show that the behavioral variety of travelers cannot be analyzed with results of one-day simulations. Development of advanced time consumption models requires observations of at least a week because daily and weekly work-leisure cycles should be included.

To expand the standard MATSim time horizon, only few changes in the current implementation have to be made. However, there are further reasons why the standard MATSim is not ideal for multi-day scenarios. First, when executing the evolutionary algorithm of MATSim, the number of possible activity sequences per agent increases exponentially with the simulation time horizon. Consequently, the process for a multi-day scenario takes too long to reach User equilibrium. Second, assigning multi-day activity sequences to every agent of a synthetic population in the beginning of the evolutionary process is a challenge due to the large number of possibilities.

In this thesis, a new approach for multi-day simulations is presented. To address the above MATSim drawbacks, a differentiation between fixed and flexible activities is proposed. An extension of MATSim to initialize agents with incomplete sequences of fixed activities and schedule on-the-fly flexible activities is implemented and tested. For flexible activity scheduling, a new multi-activity scheduler is designed. It is based on the discretization of scheduling dimensions. It solves for the number of activities, the types,
the start times and durations, the locations and the mode of transportation of trips between consecutive activities.

To evaluate the approach with real data, weekly patterns of fixed activities from frequent public transport users in Singapore are identified from smart card data and a household travel survey. With these results, *weekly fixed activity skeletons* are assigned to agents of the Singapore synthetic population. Furthermore, flexible activity type and place type models are estimated using the house interview travel survey. These results are employed to assign an *activity agenda* and a *set of evoked places* to each agent in the synthetic population.

Finally, a weekly simulation of a 10% sample of the synthetic population is executed. After only 20 iterations, 371,996 agents managed to perform all the planned activities, demonstrating the efficiency of the new approach. Results also show accurate time distributions of fixed and flexible activities. Scheduled flexible activity locations are close to reality for the majority of activity types. Travel times from/to evoked places reproduce observed distributions. Moreover, reaching *User equilibrium* is computationally tractable since 100 iterations of the weekly simulation took 2 days and 3 hours using 60 GB of RAM.
I modelli di trasporto multiagente sono metodi di supporto per la pianificazione. Per mezzo di questi approcci, futuri scenari possono essere valutati e confrontati. Con lo sviluppo di nuove tecnologie informative e della potenza computazionale, è ora possibile simulare il trasporto di $10^7$ agenti. MATSim è un modello di simulazione di trasporti multiagente su ampia scala basato su attività quotidiane. Con MATSim si può simulare in parallelo la mobilità di milioni di agenti con un ragionevole sforzo computazionale. Ad ogni modo poiché le simulazioni di MATSim sono di un solo giorno, patterns di mobilità estesi a più giorni non possono essere studiati. Questo limita le analisi su vari casi studio nell’àmbito della pianificazione dei trasporti urbani. Per esempio, recenti studi mostrano che la varietà comportamentale dei viaggiatori non può essere analizzata con risultati derivanti da simulazioni di un unico giorno. Lo sviluppo di modelli avanzati di gestione del tempo richiede osservazioni di almeno una settimana, poiché i cicli di lavoro e del tempo libero in tempi settimanali e giornalieri dovrebbero essere inclusi nelle simulazioni.

Per estendere l’orizzonte temporale standard di MATSim, sono necessari appena pochi cambiamenti nella attuale implementazione. Tuttavia, ci sono ulteriori ragioni perché l’implementazione standard di MATSim non è ideale per scenario di più giorni. In primo luogo, quando si esegue l’algoritmo evolutivo di MATSim, il numero di possibili sequenze di attività per agente aumenta esponenzialmente con il tempo di simulazione. Di conseguenza il processo simulazione di uno scenario di più giorni impiega troppo tempo per raggiungere l’equilibrio di Nash. In secondo luogo, assegnare sequenze di attività di più giorni a ogni agente di una popolazione sintetica all’inizio di un processo evolutivo è una sfida per il grande numero di possibilità.

In questa tesi, un nuovo approccio per simulazioni di più giorni è presentato. Per risolvere i precedentemente citati svantaggi, si propone una differenziazione tra attività fisse e flessibili. Un’estensione di MATSim è
implementata e testata per inizializzare gli agenti con sequenze incompleti di attività fisse e programmare all’istante attività flessibili. Per la programmazione di attività flessibili, un nuovo pianificatore di più attività è ideato, il quale è basato su discretizzazione delle dimensioni di pianificazione e trova soluzioni per il numero di attività, le tipologie, il tempo di inizio, la durata, la locazione e il modo di trasporto tra attività consecutive. Per valutare l’approccio con dati reali, pattern settimanali di attività fisse sono stati identificati dagli utenti abituali dei trasporti pubblici a Singapore per mezzo delle tessere dei trasporti e di sondaggi sui viaggi domiciliari.

Con questi risultati, l’ossatura di attività fisse settimanali sono assegnate agli utenti della popolazione sintetica di Singapore. Inoltre, modelli di attività flessibili e modelli del tipo di locazione sono stimati usando i sondaggi sui viaggi domiciliari. Questi risultati sono impiegati per assegnare un’agenda delle attività e un set di luoghi prestabili per ogni agente della popolazione sintetica.

Infine, una simulazione settimanale è valutata su un campione del 10% della popolazione sintetica. Solo dopo 20 iterazioni, 371,996 agenti sono riusciti a compiere tutte le attività pianificate, dimostrando così l’efficienza del nuovo approccio. I risultati mostrano inoltre accurate distribuzioni temporali delle attività fisse e flessibili. Locazioni pianificate delle attività flessibili sono vicine alla realtà per la maggior parte dei tipi di attività. I tempi di viaggio da/a luoghi prestabili riproducono le distribuzioni osservate. Inoltre, la valutazione dell’equilibrio di Nash è trattabile poiché il tempo di computazione di 100 iterazioni delle simulazioni settimanali è di 2 giorni e 3 ore usando 60 GB di RAM.
Acknowledgments

I started this trip leaving my beloved Bogotá in the beginning of 2011. After almost 6 years in Singapore I am finally finishing this challenging project which made me suffer and laugh so many times. First of all, I would want to thank my parents Arturo and Patricia, and my brothers Santiago and Camilo for always supporting me and making me feel loved and missed at the other side of the planet. *Los extraño mucho.* I also feel deep gratitude to my wife who accompanied me halfway; she encouraged me and helped me with my bad English. ¡Gracias hermosa!

Thanks to my advisor Professor Axhausen who was always there, but giving me freedom to research and develop my weird ideas. By the way, although he was the most of the time in Zurich, I never felt he was distant. His knowledge in every topic and his continuous desire for more were an inspiration for me. It was always a honor being supervised by Professor Axhausen in every transportation event in the world.

I also want to thank my co-examiners Professor Marcela Munizaga and Doctor Alexander Erath. Your knowledge in transportation data and constructive critics made my research improve and made me developed new abilities in Discrete choice models and Statistics.

I want to include a special recognition for my colleagues in the Future Cities Laboratory. To Pieter for his outstanding brain with which you can have deep discussions, but with a refreshing sense of humor. To Artem and Michael for their optimism, creativity and friendship. To Lijun and Teresa for teaching me so many things about different cultures, and to my new friends Ylenia, Mohsen, Temo, Sebastian and Tanvi for supporting me in the last steps.

Thanks MATSim community! Starting from the Professor Kai Nagel. Thank you for teaching me so many good programming habits and making my character mature. Thanks to Michael, Marcel, Christoph, Francesco, Michael, Thibaut, Johan and all of you. Special thanks to Milos, Maxim, Basil for the good times!
Finally I want to acknowledge the rest of my family and friends in Colombia and Singapore. My brothers Juan and Pipe, my uncles Humberto, Alfredo, William, Jorge, my aunts Sara, Clara, Elsa Victoria and Tulia and all my close cousins. Thank you to make me feel supported and cheering me up all the time with your messages.
Chapter 1

Introduction

1.1 Motivation

Multi-agent transport simulation is a tool to support decision making. Transport infrastructure projects are usually very demanding in terms of costs and time. Definition or changes in transport policies affects daily lives of millions of people in cities, regions or nations. Simulations allow to perform evaluations and comparisons of several scenarios according for different transport indicators. With the development of information technologies and computational power, it is now possible to simulate the transport system of a whole population in order of magnitude of $10^7$ agents. The activity-based approach allows to model dynamic individual transport demand with activities characterized both by type and spatio-temporal patterns. By bringing transport supply and demand together in a simulation environment, a realistic steady state solution can be obtained.

MATSim is a large-scale multi-agent transport simulation model based on activities. It can simulate the mobility of each person in a region by managing millions of agents with reasonable computation times. However, it is not designed to execute simulations of more than one day. This restricts the study of several transportation planning issues. The limitation to 24-hours scenarios also hinders the examination of spare time activities since many of them need longer periods to occur (Märki, 2014).

By expanding the time horizon to several days, comparisons of different scenarios can be performed at this scope. The decisions are not only based on indicators of an average day, but also on weekends and other periods predominantly characterized by leisure activities. Recent studies show that the behavioral variety of travelers cannot be analyzed with results of one-day simulations. Usual analytic procedures, like clustering the population based...
on travel patterns, need multi-day information to account for intrapersonal variability (Schlich, 2004). Furthermore, longer time horizons allow to include restrictions like time and money budgets, and to simulate individual mode choice over time, identifying collective preferences (Kühnimhof and Gringmuth, 2009). The development of advanced time consumption travel models require observations of at least a week for calibration purposes, because a complete cycle of work and leisure must be included (Jara-Díaz et al., 2008). Hence, the ability to simulate longer work-leisure cycles improves decision making support of transport models.

In the field of large-scale multi-agent transport simulation, this thesis contributes in the following ways:

• It will provide a method to generate realistic weekly activity plans, using classic and new data sources.
• It will provide a platform to simulate large-scale urban transport for a week time horizon and compare scenarios according to weekly indicators.
• It will provide simulation results that allow to study multi-day phenomena like routines, encounters, time and money budgets among others.

1.2 Research goal

MATSim is a toolbox to execute and analize multi-agent transport simulations based in human activities. The software is designed to manage millions of agents that interact within a common and limited transport supply. Each person in the region of interest is modelled as an agent with socio-demographic characteristics and a planned activity schedule.

As the current time horizon of MATSim is a single day, the initial plans are defined for one day; the total time of the mobility simulation is one day; public transport, road pricing or traffic lights systems are specified for one day; scoring functions are designed for one day; and re-planning strategies are designed to mutate one-day plans. Such as MATSim, other activity-based travel demand models are mostly designed for daily demand. Time use and urban transport studies also have one day time horizons. The majority of travel diaries or households travel surveys are designed to cover this period of time. However, within the study of current transport planning
challenges, daily analyses are very restrictive.

The aim of this thesis is to investigate how MATSim can be expanded to simulate and optimize multi-day transport demand, and to implement it. On the one hand, adaptations of current MATSim modules must be done to manage multi-day information. On the other hand, computational issues need to be resolved to run simulations and optimization processes within acceptable times and memory consumption. Thus, a key point of the project is to decide if the iterative optimization strategy will be retained given its high computational demand. For that reason an approach in which agents start the simulation with incomplete plans of fixed activities and schedule flexible activities on-the-fly will be implemented and evaluated against the standard MATSim. For this purpose, a weekly model of Singapore will be prepared using real data.

1.3 Structure of the document

The proposed multi-day simulation model is presented in six chapters. In Chapter 2, the activity-based approach for transport planning is introduced, followed by the description of the current implementation of MATSim. Then, a detailed explanation of the multi-activity scheduling problem and a summary of the MATSim scenario developed are presented. The chapter ends with an overview of the proposed multi-day model. Chapter 3 begins with the current situation of MATSim for multi-day scenarios and continues with a full description of an extension which allows agents to start the simulation with incomplete plans of fixed activities and schedule on-the-fly flexible activities. Results of initial tests of the new approach are included at the end of this chapter. In Chapter 4, a new recursive algorithm for multi-activity scheduling is explained in detail, and tests to evaluate the computational feasibility when processing massive scheduling tasks are carried out. New methods to recognize fixed activity weekly patterns from real data are described in Chapter 5. This includes models to detect the type of activity from public transport smart card and a clustering algorithm to extract the most popular temporal behaviors. Following, in Chapter 6, models to calculate flexible activity preferences and to model destination choices are explained and validated. Finally, all these models are applied to prepare and execute a weekly mobility simulation of Singapore. Results
Chapter 1. Introduction

and validations are presented in Chapter 7. The document also includes a final chapter to summarize the achievements, discuss the drawbacks and give guidelines for future work in this topic.

1.4 Declaration of contributions

The following Table summarizes the contributions of the author of this thesis in previously published papers.

Table 1.1: Contributions of the author in previously published papers

<table>
<thead>
<tr>
<th>Authors</th>
<th>Reference</th>
<th>Contribution</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordóñez Medina, S. A. (2016)</td>
<td>Impact of accessibility indices on secondary activity type and location type prediction using random forest classifiers, paper presented at The 5th symposium of the European Association for Research in Transportation (hEART), Delft, September 2016.</td>
<td>Calculate different accessibility indices, implement models to evaluate impacts, do the presentation and write the paper.</td>
<td>7</td>
</tr>
<tr>
<td>Ordóñez Medina, S. A. (2016)</td>
<td>Inferring weekly primary activity patterns using public transport smart card data and a household travel survey, paper presented at The 14th World Conference on Transport Research Society (WCTRS), Shanghai, July 2016.</td>
<td>Activity purpose detection, weekly primary activity patterns extraction, clustering of most popular primary activity patterns, do the presentation and write the paper.</td>
<td>5</td>
</tr>
<tr>
<td>Ordóñez Medina, S. A. (2015)</td>
<td>Recognizing personalized flexible activity patterns, paper presented at The 14th International Conference on Travel Behaviour Research (IATBR), Windsor, July 2015</td>
<td>Estimate models to assign realistic Flexible activity agendas and Sets of Evoked places, apply to the Singapore data, do the presentation and write the paper.</td>
<td>6</td>
</tr>
</tbody>
</table>
Chapter 2

Overview

2.1 Activity-based transport models and agent-based mobility simulations

The four step process (Sheffi 1985, Ortúzar and Willumsen 2011) is the traditional method of transport planning. It finds flows on transportation networks. More specifically, for an static scenario, it solves the transport economic problem of assigning demand (flows) to a given limited supply (networks). Hence, it is an static aggregated transport demand model. The models included in each step are presented below:

1. Trip generation: Given a partition of the mobility space in zones, these models estimate the number of trips produced and attracted from/to each zone. They are mainly based on aggregated socio-demographic data.

2. Trip distribution: In this step, the number of connecting trips is estimated for each origin zone - destination zone pair. These models use the number of productions and attraction estimated in the previous step, and estimated travel costs from every origin zone to every destination zone. The result is called an origin-destination matrix (OD matrix).

3. Mode choice: The share of trips using each offered transportation mode is estimated for each cell in the OD matrix. Discrete choice models are commonly used in this step. Thus, socio-economical and geographical variables determine these fractions.

4. Traffic Assignment: Given a transportation network which connects an origin zone to a destination zone in more than one way (path), the number of trips using each of these paths is calculated. This
calculation takes into account that parts of the paths (links) connecting different origin-destination zones are shared and can be congested. Travel costs for each link in the network must be provided. Two definitions of the problem were proposed: (i) **System optimum** in which the best possible assignment is found and (ii) **User equilibrium** in which it is assumed that each person traveling is greedy and intends to minimize his/her own travel time. This problem was originally formulated by Wardrop (1952).

As productions and attractions were calculated for a specific time of the day (normally during peak hours), static aggregated analyses can be performed with the link flows found by this process. When time-dependent analyses are required, the **four step process** can be modeled and executed for several periods of a day. However, with this strategy, relations between consecutive processes can not be included. Because of this, the current way to carry out dynamic studies with aggregated models is solving the Dynamic traffic assignment (DTA) problem. In recent years, transportation researchers investigate how to make DTA realistic, efficient and computationally tractable (Kaufman et al., 1998, Kawabata, 1998, Tampère et al., 2010).

Arguably, the main drawback of the **four step process** is that the congestion in links estimated in the **traffic assignment** is not taken into account in the **trip distribution**. For that reason, the four steps can be executed in an iterative way by using the results of the **traffic assignment** to re-estimate the **trip distribution**.

Dis-aggregated models have been developed in recent years as a response to the assumptions and simplifications of the **four step process**. In these models, a virtual population represents people from an area of interest. This population is generated using census data in a way that the socio-demographic structure is maintained. The mobility of each each person in the virtual population is modeled for transport planning analyses.

In **activity-based models**, the objective is to find which sequence of activities each person of the virtual population performs. The main assumption in **activity-based transport models** is that people travel because they want to perform different activities at different locations and times. Thus, the mobility demand of the population can be estimated by specifying the time and location of these activities, i.e. by estimating the start time, duration and activity facility of each activity. Thereby, assigning a sequence
of fully specified activities to each person in the virtual population replaces 
the first two steps of the four step process.

The next step in activity-based transport models is to perform a mobility 
simulation on a given transport supply i.e. number of private vehicles, paths, 
rails, roads, rails, public transport services, private transportation services, etc. 
In mobility simulations, each person of the virtual population interacts and 
competes for the transport supply while maximizing his/her utility during 
the simulation. The way of maximizing the utility is normally maximizing 
activity performing time and minimize travel time.

As transport supply is limited, the total utility of the population converges 
when every person tries to find an optimal way to travel. As every person is 
greedy when minimizing their individual utilities, this solution is a user 
equilibrium. The system optimum can also be found by making individual 
utilities depend on the effect caused to others. These solutions find the 
mode and route of every trip of every person of the population and replace 
the third and fourth step of the classic approach. Thus, this approach is a 
dynamic dis-aggregated transport demand model and detailed analyses can 
be performed at any levels.

2.2 MATSim framework

MATSim (Horni et al. 2016, Balmer 2007) is a software to simulate transport 
demand and supply interactions. Millions of agents represent the population 
of a city or region. Each agent has a transport demand represented by a 
chain of activities it has to perform during one day at different times in 
different places. All these agents are included in a mesoscopic mobility 
simulation based on queues. They interact in capacity limited transport 
networks generating dynamic congestion.

The decisions on how to travel between places to perform activities are 
made before the mobility simulation and stored as a plan for every agent. 
Thus, a plan is a sequence of activities and trips. Models to decide the 
route, start time, mode and/or destination of the journeys are included in 
MATSim and they can depend on socio-demographic characteristics.

After one execution of the mobility simulation, the experience of each 
agent is scored using a utility function. Performing activities is scored 
positively and traveling is scored negatively. In a process called Replanning,
Figure 2.1: MATSim evolutionary algorithm.

![MATSim evolutionary algorithm diagram](image)

A sample of agents is selected and their plans are modified. Activity-trip plans can be modified in many dimensions such as time, location, transportation mode or routes. A new mobility simulation is executed including the agents with modified plans. Agents are scored and the Replanning process starts again. This is repeated hundreds of times, while agents store good plans and forget bad plans. The process reaches a user equilibrium when the total score can not be improved by making individual changes on agents’ plans. This iterative process is the MATSim cycle and it is represented in Figure 2.1. This algorithm is evolutionary where the mobility simulation is a competition environment, agents are individual who compete and mutate, and activity-trip plans are their genome.

### 2.3 Multi-Activity scheduling problem

In activity-based transport models, the multi-activity scheduling problem consists of finding an optimal activity-trip sequence with a set of restrictions. The search space of this optimization problem has several dimensions, the most commonly included are: activity time, location, number of activities, activity types, and activity order. When the problem includes the transportation between activity locations, dimensions as transportation mode, route or parking location are included. Restriction are commonly
related to space (how far activity locations are) and time (how much time I have available to travel and perform activities). Monetary and social restrictions are more recently being included.

Due to the number, type and size of the dimensions the problem can not be solved as a continuous optimization problem (some dimensions are discrete), and not as a combinatorial optimization problem (the number of combinations is not tractable and some dimensions are continuous). As mentioned by Feil (2010) the fundamental problem of activity scheduling is its combinatorial complexity due to its number of dimensions. The modeler needs to develop solutions which allow solving the problem in tractable time.

Realistic human scheduling decisions are even more complex because of the quantity and variety of information involved. These decisions depend on intrinsic variables related to the behavior of the person who is planning, and on extrinsic variables related to the state of the rest of the world where the decisions are taken. Furthermore, as the decision maker doesn’t have full information, actual activity-trip sequences are not optimal.

Complex methods and algorithms have been proposed to model these multi-activity scheduling decisions with spatio-temporal restrictions such as Arentze and Timmermans (2000), Doherty et al. (2002), Miller and Roorda (2003), Arentze and Timmermans (2004), Habib (2015) or Feil (2010). Other methods are based on addressing only the next activity (Kühnimhof and Gringmuth 2009, Arentze and Timmermans 2009), or performing continuous planning without taking into account activity locations, e.g. 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), Kühnimhof 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.

The specific multi-activity scheduling problem in this thesis consists of finding the activity-trip sequence with the maximum utility given a certain time (day of the week and time of the day), a starting location, a time budget
and a future location to reach before the time budget ends. It reflects the type of problem a person encounters when, for example, he/she has some free time after work before going to home, or some free time on a Sunday without any work obligation. Hereby, time budgets are restricted to be less than one day.

In this problem the objective function is defined by the utility of performing activities and the disutility of traveling. Activity utility depends on previously performed activities, and on future planned activities. The restrictions in space are defined by the origin and the destination, while the restrictions on time by the start time, the time budget and opening times of the activity locations.

The scheduling dimensions to be solved are: number of activities, activity type, start time, duration, location and mode of transportation of trips between activities. This work aims to find a multi-activity scheduling method which doesn’t prioritize or fix any scheduling dimension.

## 2.4 Singapore scenario

The MATSim Singapore scenario was developed by Erath et al. (2012) using a Household Travel Survey (HITS 2008); public transport smart card data (CEPAS 2011); aggregated demographic data (SingStat); a navigation network (Navteq); public transport schedule in the Google format (GTFS); residential, educational, work and other facility databases; building footprints (SLA); and special trips matrices.

Figure 2.2 summarizes the development of the travel demand. It starts with a population synthesis, that is the creation of agents representing people in the study area. Although single persons do not necessarily need to be represented by a corresponding agent with the same socio-demographic attributes, the general statistical properties and correlation structure of the synthetic population should be representative for real population. Lacking a full census for the case of Singapore, an iterative proportional fitting approach on a household level was employed. Approximately 4.3 million agents were generated using controls on the household and person level from the SingStat data.

Using collected residential facilities information, households of the synthetic populations were systematically assigned to residential units
according to their type. Then, as there is not business census in Singapore, workplace capacities were estimated using CEPAS data as explained by Ordóñez Medina and Erath (2013). These results and a gravity model were employed to assigned work locations to worker agents of the population. Similarly, educational facilities were assigned to students.

Furthermore, multinomial logistic models were estimated to assign license ownership and car ownership to the agents. They contain variables regarding gender, dwelling type, household income, number of children, number of adults, etc. To capture the spatial structure of Singapore, a series of parameters indicating the postal sector of the household were estimated as well.

As mentioned previously, the travel demand generation in activity-based models consists of the assignment of activity chains to each agent in a synthetic population. After applying the models presented before, a travel decision model and an activity chain assignment model determined the local travel demand. Special trip matrices were converted to agents with short plans to model freight and international trips.

On the supply side, roads for motorized vehicles were modelled using a navigation network from NAVTeq with capacities calibrated with a planning network. More than 60,000 links and 40,000 nodes were used to model every road of the city. Locations of public transport stop and schedules were extracted from GTFS files, while public vehicles information is publicly available. A total of 4,783 stops, 331 bus lines and 10 rail lines were modeled.

To reduce computation intensity, a 25% sample of the population was simulated. Road and public transport vehicle capacities were also reduced to the same fraction. When running the simulation in a computer with 12 multi-threaded cores and 144 GB of RAM, the average computation time per iteration was 80 minutes. Within this time, the mobility simulation took 25 minutes on average, while replanning took 55 minutes. Four threads were assigned to the mobility simulation and 24 threads to replanning processes. To allocate one plans per agent, 12.2 GB of RAM were needed. The full functional scenario needed almost 100 GB of RAM, allocating 5 plans per agent and saving dynamic travel times for private and public transport. More details can be found in Erath et al. (2012).
2.5 Proposed multi-day simulation

To expand the standard MATSim time horizon (30 hours), just few changes in the current implementation have to be made. However, there are further reasons why the standard MATSim is not ideal for multi-day scenarios. First, the evolutionary algorithm of MATSim takes too long to reach user equilibrium when simulating multi-day periods, because of the combinatorial increase in the number of possible activity plans per agent to test. Second, MATSim is not ideal for multi-day simulations because of the difficulty to synthesize multi-day plans to start the evolutionary process from commonly available data. Techniques used to generate daily activity plans for a population are not efficient for plans with longer durations.

In this thesis, a new approach for multi-day simulations is presented. To address the above MATSim drawbacks, a differentiation between fixed activities and flexible activities is proposed. An extension of MATSim to...
initialize agents with incomplete plans of fixed activities and to schedule on-the-fly flexible activities is implemented and tested. With this extension, MATSim users will be able to prepare realistic multi-day activity-based demand for a synthetic population by means of data usually collected.

For flexible activity scheduling, a new multi-activity scheduler is designed. It is based on the discretization of scheduling dimensions. It solves for the number of activities, the activity types, the start time and duration, the activity locations and the mode of transportation of trips between consecutive activities.

To evaluate the method with real data, weekly patterns of primary activities of frequent public transport users in Singapore are recognized from public transport smart card data and a household travel survey. With these results, *weekly fixed activity skeletons* are assigned to agents of the Singapore scenario. Furthermore, flexible activity type and place type models are estimated using the national household interview travel survey. These results are employed to assign an *activity agenda* and a *set of evoked places* to each agent in the synthetic population. With these elements, a large-scale mobility simulation with a one week time horizon can be executed.

Next chapters describe the model in detail and present the results.
Chapter 3

Multi-agent transport simulation for a multi-day time horizon

3.1 Standard MATSim multi-day simulation

As presented in the previous chapter, MATSim was originally designed for one day mobility simulations. However, many components are capable of executing longer periods of time. In this section, details of how to run a trivial multi-day simulation using the standard MATSim are presented. The first part focuses on the transport demand, the second on the transport supply, and independent MATSim modules are reviewed in the end.

3.1.1 Multi-day transport demand in MATSim

To generate a realistic multi-day demand is still an open issue, because commonly used daily plans techniques are not efficient for longer periods of time (Bayarma et al. 2007, Arentze and Timmermans 2009, Kuhnheim and Gringmuth 2009, Feil 2010, Nijland et al. 2012, Arentze et al. 2013, Märki et al. 2014). Furthermore, survey data for periods of time longer than one day is seldom available (Munizaga et al. 2011, Stopher et al. 2008), although several efforts have been made to obtain multi-day records recently (Axhausen et al. 2002, Doherty et al. 2001, Doherty and Miller 2000, Bhat et al. 2004, Du and Aultman-Hall 2007, Bohte and Maat 2009, Löchl et al. 2005). Most of the models in this thesis are designed to solve these drawbacks.
In MATSim, the demand is defined by detailed activity plans for every person in a synthetic population. Hence, multi-day activity plans must be generated to model a realistic multi-day demand. For instance, for a weekly simulation, every plan must be defined for 168 hours or more. MATSim is able to read any time in the hh:mm:ss format and convert it to seconds (e.g. the last second of the last day of the week corresponds to the time 167:59:59 or 604,799 seconds).

From the computation perspective, to handle a multi-day demand evidently requires more RAM and computation time compared to a daily demand. As mentioned in Chapter 2, a 25% sample of the Singaporean synthetic population needs 12.2 GB of RAM to allocate one plan for more than 500,000 agents. As the size of daily plans of different days in one week is relatively uniform, more than 85 GB are needed to allocate weekly plans. Besides, in the standard MATSim process each agent retains several executed plans in memory and the default number of plans is 5. Therefore, a functional 25% sample for a weekly simulation would need approximately 430 GB.

More computation time is needed to simulate a multi-day transport demand in MATSim. Evidently, the mobility simulation takes 7 times longer than for a daily demand. For instance, one mobility simulation of the 25% Singapore scenario takes about 25 mins as described by Erath et al. (2012), therefore, a weekly mobility simulation takes almost 3 hours. Re-planning processes also take 7 times longer because weekly plans require 7 times more mutations. As a weekly simulation generates 7 times more events, computation time of Scoring and Analysis MATSim modules increase linearly with the number of simulated days. In sub-section 3.1.3, details of these modules are discussed.

3.1.2 Multi-day transport supply in MATSim

3.1.2.1 Road Network

As the MATSim road network model is not dynamic, the road network used for daily simulations works just fine for multi-day simulations. MATSim network interfaces provide methods for dynamic properties (e.g. time-dependent free speed of a link), but the standard implementations are static. Additionally, a dynamic implementation of a road network can be designed to save dynamic properties for longer periods of time.
The travel times data structure, in which dynamic travel times are recorded every iteration for each link in the network, is hard-coded with a maximum time of 30 hours. By creating a new `TravelTimeCalculatorProvider`, the total duration of the simulation can be included as the maximum time to record travel times.

### 3.1.2.2 Public Transport

The public transport model in MATSim is designed to run for longer periods of time. The transit schedule model defines times when public transport vehicles reach the stations as offsets from the departure time. Thus, to specify a multi-day public transport schedule, it is necessary to specify two aspects: (i) Times when public transport vehicles depart within multi-day intervals, i.e. 0:00:00 - 168:00:00 for a week, and (ii) the specific vehicles which are departing during different days. In reality, the same vehicle performs several departures during different days, or even within a day. If this information is available, MATSim simulates delayed departures, i.e. when a vehicle arrives late after finishing a public transport journey, its new departure will start later than the scheduled time.

### 3.1.3 MATSim modules in a multi-day simulation

#### 3.1.3.1 QSim

The default MATSim mobility simulator is designed to run simulations of cars and public transport of any duration. For a weekly simulation, it is necessary to define the end time of the simulation as 168:00:00 in the `QSim config group`.

#### 3.1.3.2 Scoring

For scoring, the `Scenario config group` includes a new parameter called `Simulation period in days` which allows to score plans with longer durations. This parameter is useful for simulation periods which are a multiple of one day, but it has two problems: (i) it is not related to the mobility simulation and (ii) if the simulation period is not a human cycle, the scoring is performed in a wrong manner. Because of the first problem, scores won’t be correct if the mobility simulation `End time` parameter and the
Simulation period in days parameter are different. The second problem happens because the default activity scoring function in MATSim called CharyparNagelActivityScoring, checks if the type of the first activity is the same than the type of the last activity and joins them as only one activity for scoring. If the simulation period is not a human cycle (e.g. the simulation starts on Tuesday and finishes on Saturday, or the simulation starts at 6 am and ends at 9 am), this cyclic assumption will score the plan wrongly. For these reasons, an extension of the CharyparNagelActivityScoring is used in these multi-day simulations. It scores plans with any duration in a non-cyclic way, but needs the start time of every plan to be specified.

In Chapter 4, another extension of the activity scoring function will be explained in detail. It takes into account that activities of the same type performed consecutively with a frequency higher than expected (e.g. shopping groceries more than twice a week) should obtain lower scores.

3.1.3.3 Re-planning

Re-planning strategies and modules are designed to mutate plans of any size within the MATSim evolutionary algorithm. However, multi-day mutations are not implemented. For example, the Sub-tour Mode Choice module selects one tour of the plan and changes the transportation mode of the trips performed in that tour. For a period of one day this mutation is substantial because an agent performs one or two tours per day. But in a week for instance, an agent can perform 14 or more tours, and changing the mode of one of those is not a substantial mutation. New re-planning modules and strategies can be designed for multi-day mutations.

3.1.3.4 Analysis

The majority of MATSim standard analyzers do not depend on the simulation duration and no modifications are needed. The Stop Watch, Travel Distance Stats, Score stats, and Trip durations modules work just fine for multi-day simulations. The only analysis module which depends on the mobility simulation duration is the Leg Histogram. The simulation end time is hard-coded as 30 hours, and the histogram is just calculated until that time. A new Leg Histogram Event handler has to be added to obtain departures and arrivals of agents for longer periods of time.
3.2 MATSim extension overview

This thesis proposes an extension of MATSim to simulate longer periods of time. To expand the standard MATSim time horizon (one day of 30 hours), few changes in the current implementation have to be made. However, there are two reasons why the standard MATSim is not ideal for multi-day scenarios. First, the evolutionary algorithm of MATSim takes too long to reach a user equilibrium with longer periods, because of the combinatorial increase in the number possible activity chains to test per agent. Second, MATSim is not ideal for multi-day simulations because to synthesize multi-day plans to start the evolutionary process from commonly available data is difficult. Techniques used to generate daily activity plans for a population are not efficient for plans with longer durations. With the proposed extension, MATSim users will be able to prepare realistic multi-day activity-based demand for a synthetic population by means of commonly available data. This section introduces the main concepts of the new strategy.

3.2.1 Fixed and flexible activities

Human activities can be classified into two categories: fixed or mandatory and flexible or optional (Arentze and Timmermans 2000; Miller 2005; Chen and Kwan 2012; Doherty et al. 2002). This differentiation is commonly used in transport science where activities or locations are classified as primary or secondary. In this thesis the two categories will be called fixed and flexible to reinforce the fact that fixed activities are prearranged.

With this classification, activity scheduling can also be divided into two steps: (i) To arrange fixed activities, (ii) To schedule flexible activities within time windows left after arranging fixed activities. Scheduling fixed activities have been extensively studied by transportation modelers and urban planners (Hansen 1959; Small 1982; Vovsha and Gupta 2013; Ordóñez Medina and Erath 2013). They have tried to find spatial patterns within geographical regions like cities, and temporal patterns during one day or longer periods of time. The challenge of the second step is the combinatorial variability of flexible activities. This makes it more difficult to find a small number of patterns that can explain flexible activity decision making. For example, when analysing the household interview travel survey
of Singapore (HITS 2012) containing daily activity chains of 1% of its population, there are more than 900 combinations of activities. By contrast, there are just 40 combinations of fixed activities after removing flexible activities. What’s more, if fixed 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. Hence, it seems a good strategy to study fixed and flexible activities separately to solve this problem.

In MATSim, activities are not classified as fixed or flexible. To start the MATSim process, it is necessary to provide fully specified activity plans for the whole population. With the extension proposed in this thesis, incomplete plans specifying only fixed activities can be handled by MATSim. When no activities are specified for an agent during a specific time window, flexible activities will be scheduled on-the-fly.

### 3.2.2 Activity scheduling on-the-fly

In reality, people plan their activities at many different times. Some activities can be planned months or years in advance while others are planned just before the activity occurs. Many flexible activity decisions are taken in this way, e.g. the decision of doing groceries after leaving the work place early. In the standard MATSim, activity plans must be fully specified, and activity scheduling decisions taken on-the-fly and under specific circumstances can not be modeled. There is an extension of MATSim called *Within-day replanning* introduced by Dobler et al. (2012). With this extension, agents start with plans containing every activity of the day, but when a special event happens, agents stop doing their planned activities and change their plans within the mobility simulation. Dobler et al. (2012) uses this extension to model exceptional events.

In this work, the proposed MATSim extension includes activity schedulers for flexible activities that are triggered while the mobility simulation is running. When an agent finds free time in its plan, MATSim will schedule new activities, times, locations and trips to reach activity locations. The concept of free time is included in the model and agents plan new activities and trips at different times during the whole mobility simulation. In Chapter 4, a utility-maximization algorithm to schedule new activities is explained in detail.
3.3 Passive planning

In this section, details of the implementation of the new MATSim extension are presented. The name of this new module is *Passive planning*, referring to the fact that agents are not active in the mobility simulation while they are scheduling new flexible activities. With this approach, a different activity scheduling optimization problem is solved on the fly at each free time window of each agent. This reduces the size of the problem which doesn’t depend anymore on the total simulation time, but on the biggest free time window. The number of iterations to reach equilibrium is reduced because agents find good local solutions for each time window. As mode choice, destination choice, route choice and time planning happen when solving each scheduling problem, it’s not necessary to include re-planning modules, at least for flexible activities. As the processes required to schedule activities are computationally expensive, a parallelization framework is also proposed and implemented.

3.3.1 Overview

The first step of the extension is to represent incomplete plans in MATSim. Plan elements in MATSim can be activities and legs. Activity times can be defined by a duration or by an end time, while leg times are defined by a travel time. An incomplete plan has at least one period of free time without any defined plan element for it. These free time windows have to be in between two activities of the plan, because no leg can be defined without a defined location.

To easily represent incomplete plans in the MATSim framework the idea is to model free time windows as a special type of Leg called *EmptyTime*. Thus, when an activity finishes while the mobility simulation is running and the next leg is an *EmptyTime*, a new module of MATSim will take care of that agent, new plan elements will be added to that plan, an the agent will be returned to the mobility simulation to continue its simulation. The travel time attribute of these new objects is used to model the size of the time window. Following, details of this approach are presented. The figures in this section are UML diagrams, a graphic language introduced by Booch et al. (2005) to represent models from the Object Oriented Programming paradigm.
3.3.2 Passive planning model

3.3.2.1 MATSim population model extension

The right side of Figure 3.1 shows the extension mentioned above: the new interface `EmptyTime` extends `Leg`. In this manner, MATSim can handle incomplete plans as sequences of activities and legs. Additionally, with the proposed approach, it is necessary to differentiate incomplete plans from executable plans. In Figure 3.1, the new interface `BasePerson` extends the standard `Person` model incorporating a new method called `getBasePlan()`. Thus, a `BasePerson` object maintains a collection of plans (person’s memory), and adds a `BasePlan` object, which is a plan with fixed `Activity` and `EmptyTime` objects defined in time, but not standard legs or flexible activities. In this way, the activities added to `BasePlan` objects define which activities are fixed within the MATSim framework.

3.3.2.2 Passive Planning agents

A simple extension of the standard MATSim agent must be implemented to integrate the proposed new processes. The new `MobsimAgent` must be aware of `EmptyTime` legs, so it doesn’t trigger leg events, and it has to
contain a Planner object as shown on the top of Figure 3.2: A Planner is a new object in charge of taking planning decisions when an agent has free time. More specifically, it replaces an EmptyTime plan element for new flexible activities and legs. It contains a collection of DecisionMaker objects which solve the different dimensions for activity and trip scheduling. The Planner object takes these results and arranges the new plan elements for the agent. On the right of Figure 3.2, the type hierarchy of the DecisionMaker is presented. This approach allows to model individual DecisionMaker objects which take individual choices based on discrete choice theory, or one object which takes all the decisions like the AgendaDecisionMaker. In Chapter 4, the algorithm used by AgendaDecisionMaker objects will be explained in detail.

### 3.3.2.3 Planning engine and Planner Manager

When an agent finishes an activity and the next plan element is an EmptyTime object, the agent will enter to a new state: PLANNING. A Planning engine object is in charge of running a series of processes for agents in this new state. Just one Planning engine object is instantiated during the whole
MATSim process. This object has three functions as shown on the top of Figure 3.3:

1. **DepartureHandler** object: In MATSim, these objects are in charge of initialize agents when they start a Leg. As EmptyTime plan elements are also legs, the Planning engine object will take any agent which initializes an EmptyTime, it will register it in an internal collection (planningAgents), and it will send the Planner object of that agent to perform the activity-trip scheduling.

2. **MobsimEngine** object: A MobsimEngine object in MATSim has a method called doSimStep(time) which is called every time step of the mobility simulation. In this method, the Planning engine reviews agents which have finished their planning process, and put them back into the mobility simulation. It also handles agents with unsuccessful planning, teleporting them to the future activity location and adding a penalty in their score.

3. **EventHandler** object: Finally the PlanningEngine works as an EventHandler, which are objects that react to specific events triggered by the mobility simulation. The PlanningEngine will collect useful information for the agents that are planning, e.g. current traffic conditions or a location’s crowdedness.

Hence, the PlanningEngine object works as a bridge between the standard mobility simulation and the new on-the-fly planning processes. However this object is not in charge of running these planning processes. The Planner object of agents in the PLANNING state are handled by a different object of type PlannerManager. This object is also instantiated once in the whole MATSim simulation. In the center of Figure 3.3 it is shown that the relationship between the PlanningEngine and the PlannerManager is 1-to-1. The reason of decoupling the planning processes from the mobility simulation is computational. Planning processes can be computationally expensive as it is shown in Chapter 4; but unlike the mobility simulation, they are fully parallelizable. Thus, while a small number of threads execute the mobility simulation, a massive number of threads can execute planning processes at the same time. The relation between the PlanningEngine and PlannerManager objects maintains the synchronicity between the mobility simulation and the planning processes. Thus, if the mobility simulation is very fast compared with the planning processes the PlanningEngine can pause the mobility simulation.
Consequently, A PlannerManager object is an administrator of planning processes. It receives Planner objects at different times, executes them, and informs the PlanningEngine object when agents are ready to continue. The current implementation of the PlannerManager has the following characteristics:

1. It creates the available number of threads for planning processes before the mobility simulations starts at every iteration. The available number of threads is the total number minus the number of threads used by the mobility simulation.

2. It rebalances the number of planning processes for each thread. In other words, it tries to maintain the same number of Planner objects in each thread while new Planner objects are received.

3. It can cancel planning processes because of their computation duration according to a specified maximum duration.

4. It informs the PlanningEngine object whether an agent is ready to continue the simulation or failed its planning process.
3.4 Application of Passive Planning

In this section, results of simulations using the new MATSim extension are presented. The first application uses Passive Planning to evaluate how useful is to have good travel time estimates when planning flexible activity locations. The second presents a multi-day scenario with an unexpected event where agents plan on-the-fly.

3.4.1 Passive Planning with travel time information

The idea with this application is to point out how useful it is for agents to know information of travel times when scheduling flexible activities. The experiment consists of running two Passive Planning processes. In the first, agents estimate the cost of traveling between two locations using the euclidean distance and average speeds. In the second, agents use updated travel times to estimate this cost. This could be implemented in reality with a digital mobile app which informs to users updated travel times between two locations.

Set-up This experiment is conducted with a 1% sample of the Singapore scenario introduced in Chapter 2 (37,425 agents). Agents start the Passive Planning process with incomplete daily plans. In a first execution, agents plan flexible activities with the activity scheduler presented in the next Chapter, using the Euclidean distance between two locations as an estimate of the travel cost. In the second, a data structure which saves updated travel times between every pair of locations is used by the agents when scheduling activity locations. Both simulations are executed in a computer with 130 GB of RAM memory and 12 cores with multi-threading. With these conditions four threads were reserved for the mobility simulations and 20 threads for Passive Planning.

Results Figure 3.4 shows the evolution of the utility through Passive Planning iterations for both executions. Updated travel times allow agents to take better activity location decisions translated into better daily utilities. The score improves 11.5% using updates travel times. This demonstrates how the Passive Planning strategy can use updated information to improve scheduling decisions.
3.4. Application of Passive Planning

Figure 3.4: Utility evolution comparison: planning flexible activities using updated travel times vs. Euclidean Distance.

3.4.2 Passive Planning: Multi-day scenario with an unexpected event

In this scenario, four similar week days (Monday to Thursday) are simulated with an unexpected event to evaluate how differently agents schedule flexible activities using Passive planning. The idea is to decrease the utility of performing shopping activities during an specific period of time and at an specific geographical area. Expected results should show how agents decide on-the-fly not to shop there.

Set-up As the first application, this experiment is conducted with a small sample of the Singapore scenario. Almost forty thousand agents (1% sample) start the simulation with four days long incomplete plans of fixed activities, generated by repeating the fixed activities found in the original daily demand. They can schedule two types of fixed activities: shop and eat.
Using the collected database of activity locations in Singapore, the shop facilities were categorized in 2 groups according to the number of stores located at the building or address. Thus, locations with more than 100 stores were categorized as Big shops and the rest as Small shops. With this information, 37 places were categorized as Big Shops and 12,983 places as Small shops in Singapore. Black pentagons in Figure 3.5 represent the geographic location of Big shops. Eating places and Small shops can be assumed to be located everywhere. The marginal utility of the flexible activities is presented in Table 3.1. These values are used in the utility maximization algorithm for activity-trip scheduling explained in next Chapter.

Table 3.1: Marginal utilities of flexible activities used in the test scenario.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Marginal utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eat</td>
<td>6 utils/h</td>
</tr>
<tr>
<td>Shop in a Big shop</td>
<td>12 utils/h</td>
</tr>
<tr>
<td>Shop in a Small shop</td>
<td>3 utils/h</td>
</tr>
</tbody>
</table>
To test the *Passive Planning* module, an unexpected event is introduced from 4pm to 9pm on Wednesday. It affects the shopping activity at facilities located inside the oval shown in Figure 3.5. To represent this event, shopping activities at these locations during the specified time period have a marginal utility of 0 utils/h. The simulation is executed in a computer with 130 GB of RAM memory and 12 cores with multi-threading. With these conditions, four threads were reserved for the mobility simulations and 20 threads for *Passive Planning*.

**Results** Figure 3.6 shows the number of trips to shop and eat planned by agents from 4pm to 9pm by day of the week. In general agents plan more trips to eat because the vast majority of shopping places are categorized as *Small shops* with a low marginal utility. Many agents prefer to eat than to travel long distances to crowded *Big shops*. The ratio between the number of eating and shopping activities changed on Wednesday comparing to the other days. When agents got the information about the event, they preferred to plan more eating activities than shopping activities on Wednesday. This change in the ratio can be appreciated more by checking the height of the lighter bars of the Figure because these bars represent the number of trips originated inside the affected area. This means that agents located inside this area changed their activity decision more than the other agents. These results demonstrate how *Passive planning* can model activity type choices.

Similarly, Figure 3.7 shows the number of trips with a shopping purpose from 4pm to 9pm for each of the simulated weekdays. On Monday, Tuesday and Thursday, agents prefer to travel to *Big shops* to obtain high utilities. In contrast, the number of trips to *Small shops* is almost the same than the number of trips to *Big shops* on Wednesday because 25 of the 37 *Big shops* are closed. As the closed facilities are located at a central location, many agents prefer to obtain a small utility at small shops, but to travel short. On Wednesday, trips to *Big shops* are planned to destinations located outside the affected area (pink bar), but unexpectedly, 23 trips are still planned to *Big shops* inside (red bar). Agents know that no activity performing utility will be obtained at these places. After analyzing these 23 cases it was found that 19 of them were trips planned to reach before 4pm, but because of traffic conditions their arrival time was at most 15 minutes later. The other 4 cases happened for a similar reason. Agents planned to reach *Big shops* in the affected area after 9pm, but they traveled faster arriving at most 6
Figure 3.6: Number of trips to shop and eat from 4pm to 9pm by day of the week.

minutes earlier. These results demonstrate how Passive planning can model activity location choices.

Figure 3.7: Number of trips to shop from 4pm to 9pm at different location types by day of the week.
Figure 3.8: Travel time distributions of trips to shop from 4pm to 9pm originated in the affected area by day of the week.

Finally, Figure 3.8 shows four travel time distributions of trips to shop originated inside the affected area. It represents how agents located in the affected area decide to travel longer to perform shopping on Wednesday during the affected period.

### 3.5 Conclusion

To simulate longer periods in MATSim is possible with few modifications. The challenges of using MATSim for longer periods of time are related to the synthesis of the activity demand. To simplify this problem, a new approach which differentiate fixed activities and flexible activities was proposed. An extension of MATSim to initialize agents with incomplete plans of fixed activities and schedule on-the-fly flexible activities was
designed implemented and tested.

The design of *Passive Planning* in MATSim is based on extensions of the population model, and the development of a new module. Two key elements in this new module are in charge of decoupling and synchronizing the mobility simulation and the planning processes: *Planning engine* and *Planner manager*. As planning processes are fully parallellizable, this approach allows to run in parallel any number of available threads for planning while the mobility simulation is executed.

Two case studies were prepared to evaluate the utility of the new approach. The first successfully shows how agents can use updated information to improve their final utility. The second shows how agents react to an unexpected event in a multi-day scenario.

The next Chapter explains in detail the algorithm to schedule activities for agents in these experiments. Following chapters use real data to create a realistic multi-day activity demand based on the concept of fixed and flexible activities.
Chapter 4

Multi activity scheduler with a time budget

The content of this chapter was previously published in Ordóñez Medina (2015a).

As explained in the introductory Chapter 2, activity scheduling is not a trivial optimization problem. It consists of finding an optimal activity-trip sequence with a set of restrictions. Due to the number, type and size of the dimensions the problem can not be solved as a continuous optimization problem (some dimensions are discrete), and not as a combinatorial optimization problem (the number of combinations is not tractable and some dimensions are continuous). As mentioned by Feil (2010) the fundamental problem of activity scheduling is its exponential complexity due to its number of dimensions (activity durations, locations, number of activities, activity types, activity order, etc.). The modeler needs to develop solutions which allow solving the problem in tractable time.

Realistic human scheduling decisions are even more complex because of the quantity and variety of information involved. These decisions depend on intrinsic variables related to the behavior of the person who is planning, and on extrinsic variables related to the state of the rest of the world when the decision is taken.

In this chapter a new method to solve the activity scheduling problem is proposed. It finds an approximately optimal activity-trip chain for a given time window according to activity performing utilities and traveling disutilities. It allows to include intrinsic variables from the decision maker and external information from the place and time where and when the decision is taken.
For weekly simulations, the algorithm will be executed to schedule flexible activities every time an agent has free time in-between two fixed activities. Assuming one free time window per agent per day on average, this algorithm must be executed 26 million times for a 3.7 million agents scenario (this is the approximate size of the full Singapore scenario presented in Chapter 2).

4.1 Overview

The specific activity scheduling problem in this thesis consists of finding the activity-trip sequence with maximum utility given a certain time (day of the week and time of the day), a starting location, a time budget and a future location to reach before the time budget ends. It reflects the type of problem a person encounters when, for example, he/she has some free time after work before going to home, or some free time on a Sunday without any work obligation.

To solve this problem, an utility maximization algorithm is design based on (i) the time geography concepts introduced by Hägerstrand (1970) and adapted by Wilson (2008), and (ii) the idea of using multi-state supernetworks for activity-trip scheduling proposed by Arentze and Timmermans (2004). In their supernetwork model, nodes in the transportation network are replicated for each possible state of a person at a given location. A state depends on the activities which were performed or will be performed, and on the state of the private vehicles of the person. Time is not part of the state. To join these nodes two types of links are modeled. Transport links change the location of the person but maintain the state. State transition links change the state but maintain the location. Besides conducting an activity, Arentze and Timmermans (2004) propose 3 more types of State transition links to manage the states they modeled: (i) leaving/returning from/to a place, (ii) parking/getting a vehicle, and (iii) picking up/dropping off people or products.

In the model proposed in this chapter, the network topology doesn’t depend on the state of the person, but on his/her possible locations at possible times from an origin to a destination restricted by a time budget. The state of a person is represented by the path, and not by a node. Hence, several paths or links can reach the same node defining a different state.
This will be explained in detail in Section 4.2.2.

In contrast to the model of Arentze and Timmermans (2004), nodes are just located at activity locations, and not in all the nodes of the transportation networks i.e. intersections and stops. There are several nodes at one location, but with different time stamps. The idea of Transport links is used to define Trip-edges when the person changes the location without performing an activity. Similarly, the idea of State transition links for conducting activities is used to define Activity-edges when the person stays at a location while time advances. The other types of State transition links introduced by Arentze and Timmermans (2004) are not modeled in this work. Furthermore, their supernetwork includes and replicate all links in the road network and in the public transport network, while in the proposed scheduler, just one link represents the trip between two locations at a specific time and by a specific mode.

In summary, this solver transforms all dimensions of the activity scheduling problem to the discrete regime and constructs a spatio-temporal graph to manage the combinations. It also uses the functions based on those introduced by Charypar and Nagel (2005) to measure activity utilities and trip disutilities. Information about car-availability and dynamic travel times can be included, as they would affect the choices of location and transportation mode.

In the following, the details of the implementation and study cases are presented.

4.2 Spatio-temporal graph based scheduling algorithm

4.2.1 Input information

The inputs include the origin \( o \) and destination \( d \), start time \( st \) and latest end time \( et \), an Activity agenda and a Set of evoked places.

As introduced in the previous chapter, an Activity agenda \( G \) is a set of \( n \) activities intended to be performed by a person. However, these activities are not defined 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 modeled as random variables,
estimating probability distributions from observations. Introduced by Axhausen (2006), 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, time windows of free time are defined in between fixed activities. The agenda can be used to plan flexible activities during these free times. Models proposed by Habib and Miller (2009) and Nijland et al. (2012) also aim to generate personalized activity agendas with an econometric and a Bayesian approach respectively. In TASHA, an activity-based model proposed by Miller and Roorda (2003), similar procedures for the proposed activity generation step are applied.

As this work also aims to resolve the locations where these flexible activities are performed, another concept must be introduced. When a person plans flexible activities, it can be assumed that he/she takes only a limited number of activity locations into consideration. Regarding the set of location 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. Thus, adapting this terminology, the limited number of locations a person considers when planning flexible activities will be called: Set of evoked places. In this thesis the term *evoked places* only refers to the locations where the decision maker is willing to travel to, in his/her decision maker mental map. Thus, a small set of *m* evoked places *P* is the second piece of input information given to this multi-activity scheduling method.

**4.2.2 Spatio-temporal graph definition**

The main algorithm consists of recursively constructing a spatio-temporal directed graph \((V, E)\) inside the space-time prism defined by the initial and final states (Hägerstrand, 1970). Each vertex is defined by a geographic location and a time stamp. Vertices are only created at evoked places.
and at specific times according to a defined time bin $\delta t$, i.e. $\forall (p, t) \in V, p \in P \land t = st + k \ast \delta t$ with $k \in \mathbb{Z}$. This controls the size of the graph. Trips and activities are represented by edges of the graph, i.e. $V = A \cup T$ where $A$ is the set of activity-edges and $T$ is the set of trip-edges. The vertices of an activity-edge must have the same geographical location ($\forall l \in A | fromVertex(l) = (pF, tF) \land toVertex(l) = (pT, tT), pF = pT \land tF < tT$) and the vertices of a trip-edge must have different geographical locations ($\forall l \in T | fromVertex(l) = (pF, tF) \land toVertex(l) = (pT, tT), pF <> pT \land tF < tT$). Each activity-edge has also an activity type or purpose associated $type(l)$, and trip-edges have a transportation mode associated $mode(l)$.

Utilities of the activity edges are calculated using an extension of the activity duration utility function presented in Charypar and Nagel (2005). Equation (4.1) presents this model:

$$U(t_d) = \beta T_x \ln \left( \frac{t_d}{T_0} \right)$$

$$T_0 = e^{\frac{-10h}{T_x \beta}} T_x$$

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 and $p$ is the priority of the activity.

This form has the following properties:

$$U(T_0) = 0$$

$$U(T_x) = \beta T_x \ln \left( \frac{T_x}{T_0} \right) = \frac{10h}{p}$$

$$\frac{\partial U}{\partial t_d}(t_d) = \frac{\beta T_x}{t_d}$$

$$\frac{\partial U}{\partial t_d}(T_x) = \beta$$

This means that when the duration of an activity is less than $T_0$, the utility is negative. When the duration is the typical duration the utility just depends on the priority of the activity. Finally, only when the duration is the typical duration, the marginal utility is corresponds to $\beta$. Thus, the
typical duration of different activities determines the optimal ratio of the durations in equilibrium.

The new version, presented in Equation 4.3 and illustrated in Figure 4.1, incurs a penalty if the same activity is performed more than once during a period of time $T$.

Because of this definition the utility of an activity-edge within an activity-trip-path depends on previous activity-edges of the path from the origin. This violates the pre-condition of the Dijkstra algorithm, where edge costs can not depend on the previous edges of the shortest path. In other words, information of the minimum cost to reach a certain vertex can not be shared among paths containing that vertex, therefore, Dijkstra algorithm cannot be applied to simplify the graph construction. Every possible path to reach a vertex has to be tested.

The travel disutility functions proposed by Charypar and Nagel (2005) are also employed for the trip-edge costs evaluating pre-calculated travel times between evoked places.

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

$$T_0 = e^{\frac{-10h}{T_x p \beta}} T_x$$

$$g(t_l) = \begin{cases} 
\alpha (e^{\gamma(T-t_l)} - 1), & \text{with } t_l < T \\
0, & \text{with } t_l \geq T 
\end{cases}$$

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

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

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, $p$ is the priority of 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 4.1), and $\gamma$ is just a function of $\alpha$. 

38
4.2. Spatio-temporal graph based scheduling algorithm

Figure 4.1: Proposed activity performing utility function.

Proposed activity utilities depend on (i) duration, and (ii) time lapsed since the last time the same activity was performed.

4.2.3 Graph construction and optimal path detection algorithm

The process starts at the initial vertex \((o, st)\), and the objective is to find the path to the final vertex \((d, et)\) with the maximum utility. From a vertex all possible activity edges and trip edges can be created. When a new vertex is created during the recursive algorithm, the path from the initial vertex is saved. When any previously created vertex is reached again through a new path from the origin, just the path with the maximum utility remains. The simplified pseudocode of this recursive algorithm is presented in Algorithm 1 and Algorithm 2.

Algorithm 1 prepares the input data, the output variable \((path)\) and calls the recursive Algorithm 2 at line 17. Algorithm 2 checks first if the current vertex is the final vertex (same time and geographic position) and checks if this complete path \((currentPath)\) has a better utility than the current maximum utility path \((path)\). If so, the maximum current path is updated at line 14. If the current vertex is not the final vertex, the graph can grow in two ways: (i) staying at the same location creating an activity-edge and (ii) moving to another location creating a trip-edge. The loop starting at line 19 creates all possible activity edges and continue the process recursively at line 27. The loop starting at line 31 creates all possible trip edges and
Chapter 4. Multi activity scheduler with a time budget

Algorithm 1 Maximum utility path

1: constants:
2: time_bin ← Fixed time bin
3: procedure getMaximumUtilityPath
4: parameters:
5: startTime ← Start time
6: startPlace ← Origin
7: endTime ← End time
8: endPlace ← Destination
9: agenda ← Activity agenda
10: places ← Set of evoked places
11: modes ← Transportation modes
12: begin:
13: global path ← null.
14: currentPath ← empty.
15: firstVertex ← createVertex(startPlace, round(startTime, time_bin)).
16: addVertexPath(currentPath, firstVertex).
17: addEdgeVertex(currentPath, round(endTime, time_bin),
18: endPlace, agenda, places, modes).
19: return path.

continue the process recursively at line 39. The Activity agenda controls the type of the activities and their duration at line 23 while the Set of evoked places controls the destinations at line 31. Another key point in the algorithm is the time discretization using the constant $\delta t$. Continuous times are rounded according to this value at the line 15 of Algorithm 1 and at lines 18 and 35 of Algorithm 2.

4.2.3.1 Algorithm complexity

The recursive algorithm reduces the problem from $d$ decisions to $d−1$ calling several times the original function. The number of operations to solve $d$ decisions is:

$$T(c, d) = c(T(d − 1))$$
$$T(c, 1) = c$$  \hspace{1cm} (4.4)

Where $c$ is the number of alternatives of each decision. Solving the
Algorithm 2 Add next edge and vertex

1: procedure \texttt{addEdgeVertex}
2: parameters:
3: \hspace{1em} \texttt{currentPath} $\leftarrow$ Path from the origin to the current vertex
4: \hspace{1em} \texttt{endTime} $\leftarrow$ End time
5: \hspace{1em} \texttt{endPlace} $\leftarrow$ Final destination
6: \hspace{1em} \texttt{agenda} $\leftarrow$ Activity agenda
7: \hspace{1em} \texttt{places} $\leftarrow$ Set of evoked places
8: \hspace{1em} \texttt{modes} $\leftarrow$ Transportation modes
9: begin:
10: \hspace{2em} \texttt{currentVertex} $\leftarrow$ lastVertex(\texttt{currentPath}).
11: \hspace{2em} \texttt{isEndTime} $\leftarrow$ currentVertex.time == \texttt{endTime}.
12: \hspace{2em} if \texttt{currentVertex.place} == \texttt{endPlace} AND \texttt{isEndTime} then
13: \hspace{3em} if utility(\texttt{currentPath}) > utility(\texttt{path}) then
14: \hspace{4em} \texttt{path} $\leftarrow$ \texttt{currentPath}.
15: \hspace{2em} else
16: \hspace{3em} \texttt{lastTravelTime} $\leftarrow$ bestTravelTime(\texttt{currentVertex.place}, \texttt{endPlace}, \texttt{modes}).
17: \hspace{3em} \texttt{time} $\leftarrow$ round(\texttt{endTime} – \texttt{lastTravelTime}, \texttt{time_bin}).
18: \hspace{2em} \hspace{2em} while \texttt{time} > currentVertex.time do
19: \hspace{3em} \hspace{3em} \texttt{vertex} $\leftarrow$ createVertex(\texttt{currentVertex.place}, \texttt{time}).
20: \hspace{3em} \hspace{3em} \texttt{duration} $\leftarrow$ \texttt{time} – currentVertex.time.
21: \hspace{3em} \hspace{3em} for \texttt{activity} $\in$ \texttt{currentVertex.place.activities} do
22: \hspace{4em} \hspace{4em} if performAct(\texttt{agenda}, \texttt{activity}, \texttt{times}) then
23: \hspace{5em} \hspace{5em} createActivityEdge(\texttt{currentVertex}, \texttt{vertex}, \texttt{activity.type}).
24: \hspace{5em} \hspace{5em} addVertexPath(\texttt{currentPath}, \texttt{vertex}).
25: \hspace{3em} \hspace{3em} addEdgeVertex(\texttt{currentPath}, \texttt{endTime}, \texttt{endPlace}, \texttt{agenda}, \texttt{places}, \texttt{modes}).
26: \hspace{3em} \hspace{3em} removeVertexPath(\texttt{currentPath}, \texttt{vertex}).
27: \hspace{3em} \hspace{3em} \texttt{time} $\leftarrow$ \texttt{time} – \texttt{time_bin}.
28: \hspace{2em} \hspace{2em} for \texttt{place} $\in$ \texttt{places} do
29: \hspace{3em} \hspace{3em} for \texttt{mode} $\in$ \texttt{modes} do
30: \hspace{4em} \hspace{4em} \texttt{travelTime} $\leftarrow$ travelTime(\texttt{currentVertex.place}, \texttt{place}, \texttt{mode}).
31: \hspace{4em} \hspace{4em} \texttt{time} $\leftarrow$ round(\texttt{currentVertex.time} + \texttt{travelTime}, \texttt{time_bin}).
32: \hspace{4em} \hspace{4em} \texttt{vertex} $\leftarrow$ createVertex(\texttt{place}, \texttt{time}).
33: \hspace{4em} \hspace{4em} createTripEdge(\texttt{currentVertex}, \texttt{vertex}, \texttt{mode}).
34: \hspace{4em} \hspace{4em} addVertexPath(\texttt{currentPath}, \texttt{vertex}).
35: \hspace{3em} \hspace{3em} addEdgeVertex(\texttt{currentPath}, \texttt{endTime}, \texttt{endPlace}, \texttt{agenda}, \texttt{places}, \texttt{modes}).
36: \hspace{3em} \hspace{3em} removeVertexPath(\texttt{currentPath}, \texttt{vertex}).
recurrence relation:

\[ T(c, d) = c^d \]  

(4.5)

In the worst case the number of alternatives of one decision, and the number of decisions during a time window are respectively:

\[ c = a \cdot p \cdot m \]
\[ d = T / b \]  

(4.6)

Where \( a \) is the number of activities in the agenda, \( p \) is the number of evoked places, \( m \) is the number of transportation modes, \( T \) is the size of the time window and \( b \) is the time bin. Replacing (4.6) in (4.5), the complexity of the problem according to the size of the input data is:

\[ T(a, p, m, T) = (a \cdot p \cdot m)^{T / b} \]  

(4.7)

This means the complexity is exponential to the time window size, and polynomial to the number of activities, evoked places and modes.

### 4.2.3.2 Graph size

In the worst case there is one vertex at every space time combination. In space just the evoked places are taken into consideration, and time is discretized according to the time bin \( \delta t \):

\[ |V|(p, T) = p \cdot (T / \delta t) \]  

(4.8)

The number of edges accounts for the number of activity-edges and trip-edges. The activity-edges cardinality depends on the number of possible activities, number of evoked places and the size of the time window, while trip-edges cardinality depends on the number of evoked places, the number
of modes and the size of the time window:

\[ |E|(a, p, m, T) = |AE| + |TE| \]
\[ |AE|(a, p, T) = a \cdot p \cdot (T/b)^2 \]
\[ |TE|(p, m, T) = p^2 \cdot m \cdot (T/b) \]
\[ |E|(a, p, m, T) = a \cdot p \cdot (T/b)^2 + p^2 \cdot m \cdot (T/b) \]  \quad (4.9)

In summary these are the relations between the size of the graph and the size on the input data:

\[ |V| \sim p, |V| \sim T \]
\[ |E| \sim a, |E| \sim p^2, |E| \sim m, |E| \sim T^2 \]  \quad (4.10)

**4.2.3.3 Algorithmic improvements**

The implemented recursive algorithm is more restrictive than the presented pseudocode. It cuts out a branch when the utility of the current path plus an expected utility does not exceed the current maximum utility. This expected utility depends on the remaining available time, and on the travel time between the current location and the final destination.

Transportation modes are also better modeled, saving the location of the car (if it was available in the beginning). When the car mode is included, but the mode of a trip edge is not car, the last car location is saved. Thus, the car mode is not available for the next locations till the saved car location is reached again. The implemented algorithm also ensures that no path has two consecutive trip edges. A maximum number of activities, or minimum activity duration can also be defined.

Thus, when the construction process finishes, the path with the maximum utility to the final vertex is found automatically. 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 duration of the activities and the places where these are performed. Optimal transportation modes of the trips between the locations are also part of the solution.
4.3 Initial tests with controlled inputs

The first way of testing the explained scheduler is comparing resulting activity-trip schedules introducing different controlled input variables. The following two study cases demonstrate the effectiveness and versatility of the algorithm.

4.3.1 Marginal utility for activity performing study case

The aim of this experiment is to investigate how the algorithm behaves when different marginal utilities for activity performing are used.

Set-up The decision maker has 10 random evoked places of two types of flexible activities (shopping and practicing sports), and an activity agenda with these two activities. The last activity performed before free time is work, finishing at 17:00, and the next scheduled activity is home starting at 23:00; i.e. the decision maker has a 6-hours time budget. The algorithm is executed 3 times with different values for the marginal utility of being at home: 6.0 utils/h, 3.0 utils/h, and 12.0 utils/h.

Results Figure 4.2 illustrates shortest paths of activities and trips found by this algorithm. In this tri-dimensional map two dimensions represent the space and the third represents the time. Evoked places are represented by colored dots on the road map. After varying the marginal utility of performing the home activity, resulting activity locations, number of activities and sequence remain, but activity durations vary according to how valuable time at home is perceived. As the home marginal utility increases, secondary activity durations are reduced proportionally to give more time to the home activity.

4.3.2 Effect of transportation mode costs on activity scheduling study case

The aim of this experiment is to investigate how the effect of transportation mode costs on activity scheduling can be represented by the activity-trip schedules calculated by the proposed algorithm. As the recursive algorithm is taking into consideration the location of the decision maker’s private...
Figure 4.2: Case study 1: Sensitivity of the optimal activity-travel path to the marginal utility of performing home activity ($\beta_{\text{home}}$).

Time window from 17:00 to 23:00.

vehicle (if he/she starts with a vehicle), the utility of a certain path can be penalized if the vehicle’s location is not the same than the decision maker’s location at the end of the time budget. This can happen when the decision maker takes different transportation modes; the simplest example consists of leaving the car at the work location, traveling by public transport, performing some other activities, and reaching home without the car.

**Set-up** The decision maker has 10 random evoked places of two types of flexible activities (shopping and practicing sports), and an activity agenda with these two activities plus a short dummy activity called visit. This activity can be performed in any location but its marginal utility is 0. The last activity performed before free time is work, finishing at 17:00, and the next scheduled activity is home starting at 23:00; i.e. the decision maker has a 6-hours time budget. The work location is intentionally located near home, and the car trips are modeled 150% more expensive than public transport trips. As explained above, the utility of an activity-trip path is penalized if the car mode is available but the vehicle does not end in the
Chapter 4. **Multi activity scheduler with a time budget**

final destination.

**Results**  Figure 4.3 shows the activity-trip path returned by the algorithm with the same visualization technique described in the previous study case. The calculated activity-trip path shows that the decision maker prefers to travel by public transport to perform activities near his/her work location (because traveling by car is expensive), return to his/her work location for the car (visit activity), and perform a short trip from there to home. The resulting path does not seem optimal because performing two trips going back home and performing a zero-utility activity at work seem utility-expensive. However, it is the best activity-trip schedule because traveling directly to home, traveling to perform secondary activities by car, or going back home without the car after traveling by public transport, are even more utility-expensive.
4.4. Using the algorithm within an agent-based simulation

Figure 4.3: Case study 2: Activity-trip chain scheduled by the proposed algorithm.

In this study case, traveling by car is 150% more expensive than traveling by public transport, the work location is located near home, and the time window is from 17:00 to 23:00.

4.4 Using the algorithm within an agent-based simulation

This section presents how the proposed scheduler is incorporated to the transport simulation platform MATSim (see also: Horni et al. (2016), Balmer (2007)). As introduced in Chapter 2, MATSim executes an evolutionary algorithm which repeats hundreds of times the mobility simulation of people
in a specific region. The simulation is agent-based. After every iteration mobility plans of a group of agents are mutated (re-planning) for the next iteration to test different initial conditions. Plans can also be modified within the mobility simulation and not in the end, this is called *within day re-planning* (Dobler, 2013). To highlight the power of multi-activity scheduling, results of the MATSim evolutionary process incorporating the proposed algorithm for *within day planning* of secondary activities are compared with results of a MATSim process using standard re-planning. The *Passive Planning* extension proposed in Chapter 3 is used to execute the presented algorithm every time an agent has free time in its plan.

**Set-up** To test the multi-activity scheduler in MATSim, a 1% sample of the full agent population (37,425) were selected, random activity agendas with 6 possible secondary activities were related (shop, eat, errands, religion, recreational and medical), and 20 evoked places were randomly assigned to each agent. The extended version of MATSim presented in the previous Chapter 3 (*Passive Planning*), allows agents to start the process with incomplete activity plans, i.e. a plan with only fixed activities specified or a skeleton as called by Doherty et al. (2002). To generate these skeletons of primary activities for the 37,425 agents, all secondary activities of their original synthesized daily plans were simply deleted. It is expected that after the first iteration (a mobility simulation of one day) every agent constructs a complete plan by filling free time windows with secondary activities by means of the proposed scheduler. After every iteration 30% of the agents are selected at random to start the next mobility simulation with incomplete plans (i.e. with the skeleton of primary activities).

To evaluate the importance of multi-activity scheduling a standard MATSim process which doesn’t change activity schedules is compared with a *Passive Planning* process starting with similar initial conditions. The complete plans generated after the first iteration of the *Passive Planning* process are used as initial plans for a standard MATSim process (normal re-planning). The same population with these plans is simulated using three re-planning strategies: Re-Route (10%), Time Allocation (10%) and Subtour Mode Choice (10%). The idea is to compare the evolution of the daily utilities executing these two strategies.

When comparing the evolution of the utilities some conditions shown in Table 4.1 must be highlighted. The standard MATSim process just mutate
4.4. Using the algorithm within an agent-based simulation

Table 4.1: Case study 3: Comparison conditions between Standard MATSim and Passive Planning.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Standard MATSim</th>
<th>Passive Planning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modules</td>
<td>ReRoute, TimeAllocation, SubtourModeChoice</td>
<td>Flexible Activities</td>
</tr>
<tr>
<td>Population mutated</td>
<td>10%, 10%, 10%</td>
<td>Scheduling</td>
</tr>
<tr>
<td>Type of mutations</td>
<td>1 Best response 2 Random routes, times, and modes</td>
<td>1 Best response with 5 dims.</td>
</tr>
<tr>
<td>Scheduling dimensions</td>
<td></td>
<td>routes, times, modes, locations and activity types</td>
</tr>
<tr>
<td>Fixed act. mutations</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Memory size</td>
<td>5 Plans</td>
<td>1 Plan</td>
</tr>
<tr>
<td>Public transport</td>
<td>Teleported</td>
<td>Teleported</td>
</tr>
<tr>
<td>Threads</td>
<td>4 mobsim 24 re-planning</td>
<td>4 mobsim 20 planning</td>
</tr>
</tbody>
</table>

three dimensions: routes, times, and modes, while the Passive Planning process mutates five: routes, times, modes, locations and activity types. The standard MATSim process can mutate durations and end times of fixed activities while the Passive Planning process can not. For the standard MATSim process agents save 5 plans in memory, for the Passive Planning just one. In both scenarios public transport trips are teleported, which means that each agent travels from an origin to the destination without interacting with others. The car trips are simulated in the road network generating congestion. The two runs are performed on the same computer. It has 130 GB of RAM memory, and 12 cores with multi-threading.

**Results** Results of these initial experiments show firstly that the proposed scheduler is computationally feasible using parallel processes in charge of several scheduling tasks. As shown in Figure 4.4, the first iteration took almost 3:30 minutes, it is the longest as every agent has an incomplete plan, while in the other iterations just 30% of them use the scheduler. The average computation time was 95 seconds.

As expected, the utility of agents increases because their activity schedules complexity increases. Lighter lines in Figure 4.5 show the evolution of the utilities after running 100 iterations of the standard MATSim, while darker lines show the evolution of the utilities using the Passive Planning MATSim. In this proposed approach, agents have fixed agendas, fixed
primary activities, and they schedule and experience different flexible activities. As only one plan is saved in each agent memory, the evolutionary algorithm consists of improving flexible activities and trips with new information through Passive Planning, and not in the selection of remembered experiences. In the final iterations, reaching user equilibrium, the executed plan score is almost always worse than the score of the plan in memory because it becomes difficult to find better solutions. For that reason, the average executed score (red line) becomes almost the same than the average worse score (green line). Additionally, scheduling events are highly heterogeneous due to the different conditions of the agents (agendas, evoked places, experienced travel times). In other words, the decisions were personalized, as every agent experienced the day in a different way (even if they live and work at the same locations).

Analyzing differences between the standard MATSim and Passive Planning in Figures 4.4 and 4.5 the following conclusions are presented:
4.4. Using the algorithm within an agent-based simulation

Figure 4.5: Evolution of the daily utilities of 37,425 agents within the MATSim evolutionary algorithm.

Lighter lines show the evolution of the utilities running 100 iterations of the standard MATSim, while darker lines show the evolution of the utilities using the Passive Planning MATSim.

- When optimizing daily experiences, the schedule of the flexible activities affects strongly the average final utility of a population. After 100 iterations the final average utility is more than 140 utils scheduling flexible activities while it is just 120 utils without this scheduling.
- Just one execution of the proposed algorithm over the entire population generates a demand close to the equilibrium in terms of routes, times and modes. The standard MATSim process just improved the utility of the initial plans from 100 to less than 120 utils. After 100 iterations mutating routes, times and modes.
- The computation of the standard MATSim process is 16% faster in the initial iterations, and 47% faster in the final iterations. This happens because the standard mobility simulation speeds up with
less congestion.

- The standard MATSim needs 66% less RAM memory (Standard MATSim: 10GB, Passive Replanning: 30GB). This figure should change with longer periods of time because more memory will be needed to store experienced plans with the standard MATSim, but the mental map with agents in Passive Planning doesn’t grow in memory.

4.5 Summary and conclusions

A recursive multi-activity scheduling algorithm that doesn’t prioritize any scheduling dimension was designed and implemented. The utility-maximization algorithm returns activity-trip chains without imposing any number of activities or trips. For each activity, the method calculates start time, duration, location and type; and for each trip the algorithm returns the travel time and the mode. Tests with controlled inputs demonstrated the potential of the model. It is possible to calibrate the number and duration of specific activities varying the corresponding utility parameters. Spatial, temporal and tour restrictions can also be imposed.

This algorithm was also tested within the activity-based multi-agent simulation platform (MATSim) to perform massive flexible activity scheduling tasks. Results show that the process is computationally feasible, the first iteration (where all agents plan secondary activities on-the-fly) took 3.5 minutes for a 1% sample (37,425 agents) running 20 threads for parallel flexible activities planning, and using 30GB of RAM memory. Furthermore, when comparing a Passive Planning process, i.e. planning secondary activities on-the-fly, with a standard MATSim evolutionary process, agents' general utility increases by more than 50%: Passive Planning improves the average utility from 100 to 140 utils. while the standard MATSim from 100 to 120 utils. This happens because agents have to schedule new secondary activities with new conditions of travel times every iteration.

For this thesis, the final goal is to use the proposed multi-activity scheduler, and the Activity agenda and Evoked places concepts to model flexible activity demand in MATSim. This will be presented in the following chapters. Longer time horizons can be simulated by synthesizing few families of fixed activities patterns, and solving secondary activities with this scheduler. Furthermore, as there is no prioritization of the scheduling dimensions,
this approach allows to model decisions such as changes in the order of activities to avoid traffic congestion, or secondary activity location choices based on the variety of activities which can be performed.

The method is fully parallelizable as the scheduling process for one person does not depend on other people. In the case of standard MATSim simulations, this process only has to be done once to prepare a realistic activity demand. As any other large scale activity scheduler, results given by this algorithm can be useful in marketing, urban planning and design, or whichever application that needs to predict when, where and why people is traveling inside a city.

If shorter computation times are needed the activity scheduling algorithm can be relaxed, assuming that activity utilities don’t depend on previous activities. In this case dynamic programming can be applied to find the optimal activity-trip chain, and the complexity would drop from $O(2^n)$ to $O(n\log(n))$, but the effect of performing the same activity type consecutively can not be modeled.
Chapter 5

Primary activity weekly patterns of public transport users in Singapore

The contents of this chapter were first published in Ordóñez Medina (2016b).

According to the approach introduced in Chapter 2, the first part of the weekly activity modeling in this thesis consists of recognizing weekly patterns of primary activities. As no reliable multi-day activity or trip survey is available in Singapore, the procedures described in this chapter are based on multi-day data extracted from public transport smart card transactions. With the introduction of smart cards for electronic payment in public transport, massive spatio-temporal data have been recorded in many cities. This data represents how public transport users from a whole city or country travel during consecutive days or weeks. It is also a very valuable data source to study travel behavior. However, as highlighted by Pelletier et al. (2011), recognizing travel or activity patterns by smart card data is a challenge because those systems were not designed to directly support planning or performance measurement.

Several studies have been developed to extract patterns from public transport transactions data. Devillaine et al. (2012) detect the activity purpose of public transport trips using smart card data of Santiago in Chile and Gatineau in Canada to compare behavioral activity patterns of users in these cities. Chakirov and Erath (2012) and Ordóñez Medina and Erath (2013) use smart card data from Singapore to detect work activities and estimate work capacities respectively. They rely on land use and transportation information to infer the type of the activity and
possible locations where users go after alighting at the public transport stops. Ordóñez Medina and Brath (2013) cluster daily work activities according to their start time and duration given a fixed number of clusters, to categorize workers. ElMahrsi et al. (2014) recognize weekly traveling patterns of public transport users clustering passengers based in temporal profiles in Rennes, France. As the authors are only interested on the travel patterns, no activity detection is proposed. However, they contextualize the clusters into socioeconomic data of the regions where public transport users board and alight. A similar approach is followed by Ma et al. (2013); they use the DBSCAN (Ester et al., 1996) clustering algorithm to identify typical daily trip chains of each individual by means of his/her temporal and geographical transactions during the course of one week. With this recognized daily trip chains, the regularity of each user is represented by a vector of 4 features. Then, they use K-means++ (Arthur and Vassilvitskii, 2007) algorithm to find clusters of similar users calculating euclidean distances between these vectors.

This chapter presents a method to identify temporal weekly patterns of primary activities performed by public transport users in Singapore. This method uses smart card transactions recorded during one week (CEPAS) and a 1% sample household travel survey (HITS). As this information is commonly available, this method is relevant in many cities. These temporal patterns are very useful for transport planners as primary activities represent the majority of people’s trips (in Singapore 79% of trips are made to perform primary activities according to HITS). Furthermore, as continuous travel data from the same user during a week can be extracted, work-leisure cycles can also be studied. Figure 5.1 displays the processes proposed in this chapter.

Activities reached and left by public transport were extracted from revealed trips of a 1% sample household travel survey, and used to develop two discrete choice models: one for workers and one for students. According to the start time and duration of an activity, these models estimate the likelihood of such activity to be HOME, WORK/STUDY or OTHER type. The first section presents details of the Singaporean travel survey, the discrete choice models’ definition and their prediction accuracy.

In the following, consistent activities were extracted from public transport smart card transactions recorded during one week. Using the mentioned discrete choice models, these extracted activities were classified by type.
Figure 5.1: Method to identify weekly activity patterns of public transport users in Singapore.

Then, for each user a 14-dimension vector summarizes his/her primary activity pattern. Each vector is composed by start time and duration of his/her work or study activities during continuous 7 days. The DBSCAN clustering algorithm was applied to these vectors in order to recognize the most common travel behaviors. The second section explains in detail these procedures, and presents the identified working and studying patterns.

In the conclusions of this chapter proposals of how the method could be extended are presented pointing out future steps. As mentioned before, this method can easily be implemented for other cities with similar data; thus, comparing results from different cities is a natural step forward.
5.1 Type of activity detection

This section describes how discrete choice models were developed to detect activity types in public transport transactions using revealed trips from the household travel survey in Singapore (HITS). The methodology presented below is based on the work presented by Chakirov and Erath (2012). They propose to estimate two independent models by splitting the travel survey observations in two: workers and students. As public transport users can also be categorized as workers or students according to the type of the smart card, these specialized models can detect the type of main activities more accurately. Activities extracted from public transport smart card data represent about 97% of public transport users in Singapore (Prakasam, 2008), during 7 consecutive days in 2012. In the following, weekly characteristics of HITS will be described in detail. Then, a summary of the travel survey information is provided, followed by a description of the implementation and evaluation of the discrete choice models.

5.1.1 Household interview travel survey summary

The household interview travel survey (HITS), carried out in 2012, asked for the daily trips of about 1% of Singapore’s citizens and permanent residents. Start time, duration, date and purpose of each trip were reported. Table 5.1 presents the number of people, trips and stages reported in HITS 2012 by day of the week.

It is assumed that during the time between two consecutive revealed trips, only one activity is performed by the respondent. Only public transport consistent activities were extracted. Public transport consistent activities must meet two conditions (see Chakirov and Erath (2012)):

- The activity location must be reached and left by public transport.
- The alighting station of the previous trip must be less than one kilometer away from the boarding station of the next trip.

With these two conditions, only activities in HITS which are from the same type than observed activities in the public transport smart card data are included in the estimation. When the first reported activity is the same than the last one, only one activity is extracted with the sum of the durations. Extracted activities are divided in two groups, student and worker activities. As the public transport system in Singapore offers special fares for students
Table 5.1: Number of observations in HITS 2012 by day of the week.

<table>
<thead>
<tr>
<th>Week day</th>
<th>Number of persons</th>
<th>Number of trips</th>
<th>Number of stages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>4,179</td>
<td>9,846</td>
<td>9,928</td>
</tr>
<tr>
<td>Tuesday</td>
<td>4,849</td>
<td>11,206</td>
<td>11,460</td>
</tr>
<tr>
<td>Wednesday</td>
<td>4,484</td>
<td>10,621</td>
<td>10,695</td>
</tr>
<tr>
<td>Thursday</td>
<td>3,915</td>
<td>9,231</td>
<td>9,607</td>
</tr>
<tr>
<td>Friday</td>
<td>3,985</td>
<td>9,235</td>
<td>10,230</td>
</tr>
<tr>
<td>Saturday</td>
<td>2,894</td>
<td>6,519</td>
<td>7,303</td>
</tr>
<tr>
<td>Sunday</td>
<td>2,348</td>
<td>5,266</td>
<td>5,639</td>
</tr>
<tr>
<td>Total</td>
<td>26,654</td>
<td>61,924</td>
<td>64,862</td>
</tr>
</tbody>
</table>

it is possible to recognize some of them according to the type of smart card they use.

HITS respondents must be categorized as workers or students in a consistent manner with public transport users recorded in the Smart card data. Table 5.2 presents the types of public transport cards used in Singapore, and how they are recorded. As is shown in this table, tertiary students can not be recognized according to the card type. The Student Smartcard is only available for students below 21 years old. On the other hand, all types of education (primary, secondary, college or polytechnic, university, etc) are included in HITS, and the age is reported in intervals of 5 years. Thus, full time students below 20 years are most probably carrying a student card, but it is uncertain for students between 20 and 25 years old. As an assumption, students between 20 and 25 years old and who reported to attend an intermediate education institution (college or polytechnic) in HITS, are assumed to carry a student card. The rest of people in HITS are categorized as workers. Tertiary education is therefore sometimes detected as a work activity by this method.

5.1.2 Discrete choice models for activity type

Public transport consistent activities in HITS were used to estimate one of two discrete choice models: Study model and Work model. The alternative activity types of the Study model are (i) HOME, (ii) STUDY and (iii)
Table 5.2: Types of smart cards used in Singapore and the way the types are recorded.

<table>
<thead>
<tr>
<th>Recorded type of user</th>
<th>Type of card</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child/Student</td>
<td>Child Concession Card</td>
</tr>
<tr>
<td></td>
<td>Student Smartcard</td>
</tr>
<tr>
<td>Adult</td>
<td>Tertiary Student Concession</td>
</tr>
<tr>
<td></td>
<td>NSF Concession Card</td>
</tr>
<tr>
<td></td>
<td>Adult Stored Value Smartcard</td>
</tr>
<tr>
<td>Senior Citizen</td>
<td>Senior Citizen Concession Card</td>
</tr>
</tbody>
</table>

*OTHER*, while the alternative activity types of the *Work model* are (i) *HOME*, (ii) *WORK* and (iii) *OTHER*. The type *OTHER* is the reference alternative in both models. To estimate how the start time and the duration determine the type of an activity, the following data analysis is carried out:

Equations (6.3) give the probability of an activity $a$ to be of type $t$:

\[
\forall t \in T
\]

\[
U_{at} = V_{at} + \epsilon_{at}
\]

\[
y_{at} = \begin{cases} 1, & \text{if type } t \text{ is reported} \\ 0, & \text{if a different type is reported} \end{cases}
\]

\[
\epsilon_{at} \sim \text{Logistic}
\]

\[
P(y_{at} = 1) = \frac{e^{V_{at}}}{\sum_{u \in T} e^{V_{au}}}
\]

Where $T$ is the set of activity types, $V_{at}$ is the observed utility of the activity $a$ when the type is $t$, $\epsilon_{at}$ is the unobserved utility, and $P(y_{at} = 1)$ is the probability that an activity $a$ is of type $t$. For the reference type $r$, $V_{ar}$ is zero to make the other utilities relative to the reference type. The objective is to find out how the duration of activity $d_a$ and its start time $s_a$ are related to the observable utilities $V_{at}$. To visualize these relations, duration and start time dimensions were binned in 15 minutes intervals.
For each interval, the number of observations of an alternative activity type is compared with the number of observations of the reference alternative (OTHER activities). To visualize the utility of an alternative compared with the reference category utility the following estimate is employed:

\[
P(y_{at} = 1) = \frac{N_{at}}{\sum_{u \in T} N_{au}} = \frac{e^{V_{at}}}{\sum_{u \in T} e^{V_{au}}} \quad (a)
\]

\[
P(y_{ar} = 1) = \frac{N_{ar}}{\sum_{u \in T} N_{au}} = \frac{e^{V_{ar}}}{\sum_{u \in T} e^{V_{au}}} \quad (b)
\]

Replacing (b) in (a) \( (5.2) \)

\[
\frac{N_{at}}{\sum_{u \in T} N_{au}} = e^{V_{at}} \frac{N_{ar}}{\sum_{u \in T} N_{au}} \quad (c)
\]

\[
\log \left( \frac{N_{at}}{N_{ar}} \right) = V_{at} \quad (d)
\]

This means that the logarithmic ratio of the number of observations of an alternative with the number of observations of the reference alternative is a good estimation of the value of the observed utility of that alternative. The bars of Figure 5.2 and 5.3 show this logarithmic ratio calculated to every interval in the duration and start time dimensions respectively, for the alternative HOME and the alternative STUDY related to the reference alternative OTHER. When the number of observations of the alternative type is 0 (the logarithmic ratio is negative infinite) the bar is drawn with the minimum height, when the number of observations of the reference type is 0 (the logarithmic ratio is positive infinite) the bar is drawn with the maximum height, and when both are 0 (the logarithmic ratio is undefined) the bar is not displayed (height 0). In the same way, Figure 5.4 and 5.5 display this analysis for the duration and start time dimensions of the Work model respectively. These bar plots help in modeling the observable utility functions of the two discrete choice models.

Red lines in Figures 5.2, 5.3, 5.4 and 5.5 show the resulting utility functions as defined in equations 5.3 and 5.4. Intervals of these linear piece-wise functions were manually defined according to the bar plot explained previously. Optimal slopes are the result of the log-likelihood minimization. When overlapping these results, it can be appreciated how
the estimated linear piece-wise functions match the presented bar plots.

\[ V_{a\text{Home}} = K_{\text{home}} + \beta_{\text{home}} d_{1-16} + \beta_{\text{home}} s_{8-18} \]
\[ V_{a\text{Study}} = K_{\text{study}} + \beta_{\text{study}} d_{1-4} + \beta_{\text{study}} d_{7-8} + \beta_{\text{study}} s_{8-14} \]

\[
d_{x-y} = \begin{cases} 
0 & d_a < x \\
(d_a - x) & x < d_a < y \\
y - x & d_a > y 
\end{cases} 
\] (5.3)

\[
s_{x-y} = \begin{cases} 
0 & s_a < x \\
(s_a - x) & x < s_a < y \\
y - x & s_a > y 
\end{cases} 
\]

\[ V_{a\text{Home}} = K_{\text{home}} + \beta_{\text{home}} d_{5-15} + \beta_{\text{home}} s_{8-21} \]
\[ V_{a\text{Work}} = K_{\text{work}} + \beta_{\text{work}} d_{1-9} + \beta_{\text{work}} s_{7-20} \]

\[
d_{x-y} = \begin{cases} 
0 & d_a < x \\
(d_a - x) & x < d_a < y \\
y - x & d_a > y 
\end{cases} 
\] (5.4)

\[
s_{x-y} = \begin{cases} 
0 & s_a < x \\
(s_a - x) & x < s_a < y \\
y - x & s_a > y 
\end{cases} 
\]

The coefficient of determination of the study model is 0.944, and its final log-likelihood is -415.237, using only 7 parameters with 6,717 observations. The coefficient of determination of the work model is 0.777, and its final log-likelihood is -4,599.333, using 6 parameters with 18,792 observations. Tables 5.3 and 5.4 summarize the significance of the parameters in the study model and the work model respectively. To test the prediction ability of the model, 10% of the observations were separated from the training set and both discrete choice models were estimated again with the rest of observations. This 10% sample was used to test the estimated models. The study model predicted correctly 87.7% of the activity types for students in the survey, while the work model predicted in the right way 82.6% of the
5.1. Type of activity detection

Figure 5.2: Logarithmic rate of the number of observations of activities of type *HOME* and type *STUDY* with the reference type *OTHER* by duration.

The more transparent the color of a bar the more observations within the 15 minutes bin.
Figure 5.3: Logarithmic rate of the number of observations of activities of type HOME and type STUDY with the reference type OTHER by start time.

The more transparent the color of a bar the more observations within the 15 minutes bin.
Figure 5.4: Logarithmic rate of the number of observations of activities of type *HOME* and type *WORK* with the reference type *OTHER* by duration.

The more transparent the color of a bar the more observations within the 15 minutes bin.
Figure 5.5: Logarithmic rate of the number of observations of activities of type \textit{HOME} and type \textit{WORK} with the reference type \textit{OTHER} by start time.

The more transparent the color of a bar the more observations within the 15 minutes bin.
5.1. Type of activity detection

Table 5.3: Resulting parameters of the study model utility functions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{home}$</td>
<td>-6.07</td>
<td>-12.94</td>
</tr>
<tr>
<td>$\beta_{homeD}$</td>
<td>0.722</td>
<td>17.90</td>
</tr>
<tr>
<td>$\beta_{homeS}$</td>
<td>0.431</td>
<td>8.67</td>
</tr>
<tr>
<td>$K_{study}$</td>
<td>1.7</td>
<td>3.14</td>
</tr>
<tr>
<td>$\beta_{studyD1}$</td>
<td>1.25</td>
<td>9.16</td>
</tr>
<tr>
<td>$\beta_{studyD2}$</td>
<td>-1.97</td>
<td>-2.71</td>
</tr>
<tr>
<td>$\beta_{studyS}$</td>
<td>-0.748</td>
<td>-11.53</td>
</tr>
</tbody>
</table>

Table 5.4: Resulting parameters of the work model utility functions.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{home}$</td>
<td>-6.41</td>
<td>-43.50</td>
</tr>
<tr>
<td>$\beta_{homeD}$</td>
<td>0.884</td>
<td>46.92</td>
</tr>
<tr>
<td>$\beta_{homeS}$</td>
<td>0.494</td>
<td>36.60</td>
</tr>
<tr>
<td>$K_{work}$</td>
<td>-2.79</td>
<td>-24.58</td>
</tr>
<tr>
<td>$\beta_{workD}$</td>
<td>0.685</td>
<td>46.16</td>
</tr>
<tr>
<td>$\beta_{workS}$</td>
<td>-0.143</td>
<td>-12.03</td>
</tr>
</tbody>
</table>

workers’ activity types from the test sample.

Because the day of the week when an activity is performed seems to be a good predictor for its type, the work model and the study model are also estimated including extra parameters to model weekends or Sundays. Equation 5.5 and 5.6 present the functional form. The same observations from the travel survey were used to find optimal parameters (6,717 observations for the study model and 18,792 observations for the work model). One parameter was added for each significant parameter of the original model to measure the effect of the weekday when the activity was performed: $\beta_{homeD}Week$, $\beta_{homeS}Week$, $\beta_{studyD1}Week$, $\beta_{studyD2}Week$, $\beta_{studyS}Week$, $\beta_{workD}Week$, $\beta_{workS}Week$. External alternative parameters were also included: $K_{homeWeek}$, $K_{studyWeek}$, $K_{workWeek}$. None of these
new parameters was significant and their estimates are close to 0 as shown in Table 5.5. This means start times and duration of working or studying activities performed by public transport users on weekends in Singapore don’t change significantly. As the external alternative parameters are also not significant, the proportion of the number of working or studying activities and the total number of activities doesn’t change either during weekdays. Although HITS shows that fewer working and studying activities are performed on weekends, these results show that the number of other activities performed by public transport users in Singapore decreases too.

\[
V_{aHome} = K_{home} + K_{homeWeek}w + \beta_{homeD}(1 + \beta_{homeDWeek}w)d_{1-16} + \\
\beta_{homeS}(1 + \beta_{homeSWeek}w)s_{8-18}
\]

\[
V_{aStudy} = K_{study} + K_{studyWeek}w + \beta_{studyD1}(1 + \beta_{studyD1Week}w)d_{1-4} + \\
\beta_{studyD2}(1 + \beta_{studyD2Week}w)d_{7-8} + \beta_{studyS}(1 + \beta_{studySWeek}w)s_{8-14}
\]

\[(5.5)\]

\[
V_{aHome} = K_{home} + K_{homeWeek}w + \beta_{homeD}(1 + \beta_{homeDWeek}w)d_{5-15} + \\
\beta_{homeS}(1 + \beta_{homeSWeek}w)s_{8-21}
\]

\[
V_{aWork} = K_{work} + K_{workWeek}w + \beta_{workD}(1 + \beta_{workDWeek}w)d_{1-9} + \\
\beta_{workS}(1 + \beta_{workSWeek}w)s_{7-20}
\]

\[(5.6)\]

Where \(w\) is a dummy variable representing if the activity was performed on a weekday.
Table 5.5: Resulting extra parameters to estimate impact of the week day on the activity type.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{\text{homeWeek}}$</td>
<td>2.46</td>
<td>1.57</td>
</tr>
<tr>
<td>$K_{\text{studyWeek}}$</td>
<td>0.83</td>
<td>0.57</td>
</tr>
<tr>
<td>$\beta_{\text{homeDWeek}}$</td>
<td>-1E-3</td>
<td>-0.01</td>
</tr>
<tr>
<td>$\beta_{\text{homeSWeek}}$</td>
<td>3E-3</td>
<td>0.04</td>
</tr>
<tr>
<td>$\beta_{\text{studyD1Week}}$</td>
<td>0.09</td>
<td>0.46</td>
</tr>
<tr>
<td>$\beta_{\text{studyD2Week}}$</td>
<td>-0.11</td>
<td>-0.07</td>
</tr>
<tr>
<td>$\beta_{\text{studySWeek}}$</td>
<td>-0.1</td>
<td>-0.9</td>
</tr>
<tr>
<td>$K_{\text{homeWeek}}$</td>
<td>0.09</td>
<td>0.18</td>
</tr>
<tr>
<td>$K_{\text{workWeek}}$</td>
<td>7E-5</td>
<td>0.0</td>
</tr>
<tr>
<td>$\beta_{\text{homeDWeek}}$</td>
<td>0.03</td>
<td>0.81</td>
</tr>
<tr>
<td>$\beta_{\text{homeSWeek}}$</td>
<td>-0.01</td>
<td>-0.64</td>
</tr>
<tr>
<td>$\beta_{\text{workDWeek}}$</td>
<td>0.01</td>
<td>0.74</td>
</tr>
<tr>
<td>$\beta_{\text{workSWeek}}$</td>
<td>-2E-3</td>
<td>-0.09</td>
</tr>
</tbody>
</table>

5.2 Weekly primary activity patterns

This section describes in detail the public transport smart card data used for this work, how activities were identified in this data, the application of the models developed in the previous section 5.1 to detect the type of the activity, the definition of personal weekly primary activity patterns and the clustering algorithm used to infer the main patterns in Singapore.

5.2.1 Extracting activities from public transport smart card data in Singapore

As mentioned before, 97% of public transport trips are paid by smart card in Singapore. Users have to tap-in when entering the system and tap-out when exiting it. For the bus system, these transactions happen when boarding and alighting the vehicle, while in the train systems (MRT and LRT) the transactions occur when reaching or leaving the stations. EZ Link, the company that runs the smart card-based fare collection system, collects
data of each transaction made in the city (CEPAS data). This company is a subsidiary of Land Transport Authority in Singapore.

This work uses data of transactions made in 7 consecutive days in 2012. As a first step, trips were extracted from consecutive tap-in and tap-out transactions. As the objective is to recognize activities of public transport users, the relevant information from each trip is: the card id, the start time, its travel time, the location of the origin, the location of the destination, if the trip was a transfer, and the type of passenger card. All the records are split in two (students and workers) depending on the type of card the passenger holds (Student Smartcard or others). Activities are extracted from consecutive trips of the same card id (or the last and the first for the last activity). In the same way as explained previously, it is assumed that just one activity is performed between two consecutive trips. As also mentioned in the section 5.1, only public transport consistent activities must be extracted. This means that the distance between the destination of the previous trip and the origin of the next trip has to be less than 1 km.

From the 3’348,628 users registered during one week, 3’026,300 users have at least one consistent activity identified, 1’049,036 users with at least one activity during five different days, and only 274,005 users with activities observed every day of the week. With 20’856,442 public transport consistent activities, an average of 6.89 consistent activities were extracted by user.

The next step is to apply the discrete choice models presented in section 5.1 to the extracted activities from the CEPAS data. To determine the type of each activity, equations 5.3 or equations 5.4 are evaluated with extracted start time and duration values (depending if the public transport user is categorized as worker or student). Then, equation 6.2 is evaluated with these utilities (and 0 utility for the OTHER type) in order to estimate the probability of each type. Finally, one of the types is selected randomly using these probabilities.

A total of 9’346,237 HOME activities, 4’111,340 work activities, 829,842 study activities and 6’569,023 other activities were recognized. Table 5.6 summarizes the number of activities by type and day in the week.
Table 5.6: Number of identified public transport consistent activities by type and day of the week.

<table>
<thead>
<tr>
<th>Day</th>
<th>HOME</th>
<th>WORK</th>
<th>STUDY</th>
<th>OTHER</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>1,399,959</td>
<td>678,637</td>
<td>129,245</td>
<td>840,970</td>
<td>3,048,811</td>
</tr>
<tr>
<td>Tue</td>
<td>1,419,886</td>
<td>691,697</td>
<td>130,553</td>
<td>867,149</td>
<td>3,109,285</td>
</tr>
<tr>
<td>Wed</td>
<td>1,428,431</td>
<td>691,526</td>
<td>131,500</td>
<td>885,047</td>
<td>3,136,504</td>
</tr>
<tr>
<td>Thur</td>
<td>1,416,830</td>
<td>689,473</td>
<td>131,543</td>
<td>884,313</td>
<td>3,122,159</td>
</tr>
<tr>
<td>Fri</td>
<td>1,437,972</td>
<td>672,698</td>
<td>143,080</td>
<td>1,032,078</td>
<td>3,285,828</td>
</tr>
<tr>
<td>Sat</td>
<td>1,216,291</td>
<td>384,457</td>
<td>104,829</td>
<td>1,100,030</td>
<td>2,805,607</td>
</tr>
<tr>
<td>Sun</td>
<td>1,026,868</td>
<td>302,852</td>
<td>59,092</td>
<td>959,436</td>
<td>2,348,248</td>
</tr>
</tbody>
</table>

5.2.2 Inference of the main weekly primary activity patterns in Singapore

A personal weekly primary activity pattern is the way a person performs primary activities during one week, e.g., a public transport user works 8 hours for five days in the week and just half time on Saturdays. A vector of 14 numbers was designed to model each weekly pattern. For each day in the week, two numbers represent the primary activity different from HOME of each user. The first number is the duration of the activity and the second is the start time. When more than one primary activity is performed the same day, the earliest start time and the sum of the durations represent the pattern. When no primary activity is extracted in one day, the duration is 0 and the start time is a negative constant (-24 hours). More than 3.3 million vectors of 14 dimensions are generated to represent the personal weekly primary activity patterns of public transport users in Singapore. Among them, 3.1 million are work patterns and 250,000 are study patterns.

The final objective of this work is to find the most common weekly primary activity patterns by clustering the vectors just described. The clustering processes are applied to public transport users with at least 6 activities (HOME, WORK or STUDY and OTHER) detected during the week. This filters transactions of users who don’t frequently use public transport. From the 3.1 million work patterns, 1.2 million have at least 6 activities detected during the week (pt-week-consistent workers); and from
the 250,000 study patterns, only 140,000 (pt-week-consistent student). As inference and clustering techniques will be applied to pt-week-consistent users, results and findings just represent this fraction of the population: 39% of workers and 56% of students who use public transport. Furthermore, activities from pt-week-consistent users which are not pt-consistent won’t be taken into consideration.

Clustering algorithms are commonly categorized in six types: (i) Centroid based, (ii) Connectivity based, (iii) Density based, (iv) Probabilistic, (v) Dimensionality Reduction, and (vi) Neural Networks/Deep Learning. Centroid-based algorithms need previous knowledge for initialization and to define the number of clusters. Connectivity based algorithms are expensive to compute \(O(N^2)\), the number of clusters must be given as a termination condition, and they are sensible to noise. Probabilistic algorithms are designed to cluster discrete variable combinations. Even if the number of dimensions of the input information is reduced dramatically (which is normally not the case) the number of clusters can be massive, based on combinations of the principal components. Finally, neural networks are used to cluster using self organized maps, but the number of clusters has also to be given and observations can not be categorized as noise.

The density-based spatial clustering of applications with noise (DBSCAN) proposed by Ester et al. (1996) is employed for these two clustering tasks. DBSCAN is an algorithm which recognizes clusters at dense regions (i.e. where many point are located in the same space). The algorithm needs two parameters as input, a search radius \(\epsilon\) and a minimum number of points to be considered a dense region \(\text{minPts}\). The algorithm starts from an arbitrary point, then it finds all the points located within an \(\epsilon\) distance \(\epsilon\_neighborhood\) and checks if the number of points is more or equal than \(\text{minPts}\). If the number of neighbors is not enough the \(\epsilon\_neighborhood\) is considered noise, otherwise it is labeled as a dense area and a cluster starts. The \(\epsilon\_neighborhood\) of every point in the cluster is retrieved, and only the points in the \(\epsilon\_neighborhoods\) labeled as dense areas, are added to the current cluster. When all points in the cluster are visited and no more points are added, the algorithm starts the process of forming another cluster from an arbitrary point which still has not been visited. The algorithm finishes when all points are visited and labeled as dense (belonging to a cluster) or noise.

The DBSCAN algorithm is suited to the activity patterns clustering tasks.
for many reasons. The algorithm doesn’t need the number of clusters to be specified, which is very convenient as the number of main weekly primary activity patterns is unknown. The search radius and the distance function can be adjusted to control weekly activity patterns to be in the same or different clusters. Finally, not all the points are clustered, some are labeled as noise, which is convenient as it is expected to find exceptional weekly primary activity patterns which shouldn’t affect the big clusters.

Equation 5.7 shows the distance function defined for the clustering task. In other words, this function determines how different two weekly patterns or two vectors are. This distance is measured in time units.

\[
D(u, v) = p(u, v) + \sum_{w \in W} \text{abs}(u_{w, du} - v_{w, du}) + \text{abs}(u_{w, st} - v_{w, st})
\]

\[
p(u, v) = \begin{cases} 
12h & \text{if } u \text{ and } v \text{ are intuitively different} \\
0 & \text{if } u \text{ and } v \text{ are not intuitively different}
\end{cases}
\]  

(5.7)

Where \( u \) and \( v \) are two weekly patterns, \( p(u, v) \) is a penalty for patterns intuitively different, \( W \) is the set of days in the week, \( u_{w, du} \) is the duration of the \( u \) primary activity in the day \( w \), and \( u_{w, st} \) is the start time of the \( v \) primary activity in the day \( w \). Patterns intuitively different are defined by the modeler; this penalty prevents these patterns to finish in the same cluster. For the Singapore model, two people have intuitively different patterns when:

- One of them works/studies at least once in the week and the other doesn’t
- In two or more weekdays, one of them works/studies and the other doesn’t
- In one or more weekends, one of them works/studies and the other doesn’t

On one hand, these conditions are more flexible for weekdays, because it is assumed that if a person traveled many weekdays to work/study, his/her primary activity has a stable schedule, his/her pattern can be then compared with a person who traveled one day more, and they could possibly end in the same cluster. On the other hand, a single difference on weekends should be penalized to find out primary activity behaviors on these days; i.e. a person who works/studies on Saturdays and Sundays is intuitively different.
to a person who works/studies just on Saturdays. It was also decided to include Friday as a weekend for the work activity clustering to allow the algorithm to show differences in working activities during this day. If working activities on Fridays differ significantly from the other days of the week, the result would include large clusters with patterns excluding working activities on Friday.

For computation issues related to the size of the distance matrix in RAM memory, a 10% sample of the vectors were used for the clustering tasks. For the work activity pattern clustering, after pre-processing the distance matrix, DBSCAN found more than 100 clusters using $\epsilon = 2h$ and $minPts = 5$. The value of the $\epsilon$ parameter means that two vectors with more than two hours of time-difference shouldn’t be in the same neighborhood, and the value of $minPts$ represents that at least 5 observations are needed to create a new neighborhood. Only 25% of the vectors were not clustered (noise). Figure 5.6 presents the median start time and duration of the 17 largest clusters ordered by size. These represent the most common weekly work activity patterns of public transport users in Singapore. As expected, one of the most common behavior is working from Monday to Friday for at least 8 hours; this is represented by the clusters 4, 24 and 76, although workers in cluster 24 finish late on Friday and people in cluster 76 always start to work at noon. Counting the total number of public transport users in these clusters, 5-weekday workers represent 45% of the population. People without recognizable work activities at all resulted the 2nd biggest cluster with 19% of the population, and in contrast, people with consistent working activities everyday is represented by smaller clusters (3 and 6), representing 4% of the population. Clusters 3 and 6 are different because in cluster 3 people start and end working earlier. The third largest individual cluster represents workers who work every weekday except Friday. This supports the inclusion of Friday as a weekend; without this penalization, public transport users without a recognizable working activity on Friday could have ended in other clusters.

The graphical representation of the clusters in Figure 5.6 allows to determine the flexibility of the working activity. For cluster 4, the rounded shape on the left of every working activity box, represents high flexibility of the working activity start time. On the other hand, for cluster 3, the left side of the shape of the seven working boxes, represents an inflexible work activity start time.
5.2. Weekly primary activity patterns

Figure 5.6: Main weekly work activity patterns for public transport users in Singapore.

Figure 5.7 shows the 17 largest clusters of the study activity vectors. DBSCAN was executed using $\epsilon = 2h$ and $minPts = 5$, with a total of 5,746 observations (38%) classified as noise. As expected, the largest cluster represents the behavior of studying for more than 8 hours during 5 weekdays (24%). The second represents a shorter study activity on Friday.
Figure 5.7: Main weekly study activity patterns for public transport users in Singapore.

Many other clusters are students who went a shorter time to the school on other days (19, 52, 11, 86). Cluster 38 represents half time students, while clusters 20, 13 and 27 represent students who also study on Saturdays. Finally, no significant cluster includes an education activity on Sunday.
5.3 Conclusions and discussion

In this chapter, temporal weekly patterns of primary activities performed by frequent public transport users in Singapore were recognized using Smart Card Data from public transport transactions. The proposed method consists of two steps. In the first, activities reached and left by public transport were extracted from a 1% sample travel survey of Singapore. Then, to estimate the likelihood of an activity to be HOME, WORK/STUDY or OTHER type by its duration and start time, the extracted activity records were used to develop two discrete choice models: one for workers and one for students. In the second step, consistent activities were extracted from public transport smart card transactions recorded during one week, and classified by type applying the estimated discrete choice models. Then, to summarize weekly primary activity patterns, 14-dimension vectors composed by start time and duration of WORK or STUDY activities during continuous 7 days, were generated for each frequent public transport user. Finally, to recognize the most common travel behaviors, the DBSCAN clustering algorithm was applied to these vectors.

Results show that 5-weekday workers are the most representative group of public transport users. These persons are represented by three large clusters with differences on the start time and duration of the recognized working activities. Users without recognized working activities during the week were represented by the 2nd largest cluster, and in contrast, users with recognized working activities everyday were represented by two smaller clusters. The third largest individual cluster represents frequent public transport users with recognized working activities every weekday except Friday. For students, some of the largest clusters represent the behavior of studying more than 8 hours during the 5 weekdays. These clusters differ due to the duration of the study activity on Friday. A cluster with 3% of the population represents students who study every weekday except on Friday. Half time students, and people studying 6 days were also represented by small clusters.

In Chapter 7, comparisons between consistent public transport users and the rest of the population temporal behaviors will be presented to evaluate if the clusters found in this work can be generalized. Furthermore, socioeconomic and geographical variables are related to the resulting groups, in order to contextualize the findings. Although no socioeconomic
characteristics of the public transport users are known, census information aggregated by zones were used to estimate socioeconomic shares, which can be interpreted as probabilities. In other words, if the home location and the work location of a certain user can be extracted from his/her public transport data, and these locations belong to zones with certain socioeconomic distributions, these distributions can be related to the user.

Additionally, similarly to the work by Ordóñez Medina and Erath (2013) and Chakirov and Erath (2012), mode share maps can be extracted from the travel survey in order to estimate weekly work or study patterns of people who don’t frequently use public transport.
Chapter 6

Personalized flexible activity scheduling using Singapore data

The contents of this chapter were first published in Ordóñez Medina (2015b).

With the results of the previous Chapter 5, it can be assumed that mandatory activities like rest, work or study, can be arranged and fixed, using recognized weekly patterns. This chapter focuses on planning or scheduling flexible activities and trips during the remaining time windows using data available for Singapore. Understanding how individuals from a certain city or region plan flexible 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). The method presented is designed to extract and model these flexible activity patterns from real Singapore data, and to use them to make activity chain predictions. As mentioned in Chapter 4, 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 such as Arentze and Timmermans (2000), Doherty et al. (2002), Miller and Roorda (2003), Arentze and Timmermans.
Chapter 6. Personalized flexible activity scheduling using Singapore data

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, e.g. 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 available data from Singapore (i.e. travel survey 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 and spatial 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, the concept introduced in Chapter 4. It 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 the same personalized information is used to estimate a Set of Evoked places, another concept introduced in Chapter 4. Finally the graph-based utility maximization algorithm explained in the previous chapter, takes the agenda, the set of evoked places, 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). In other words, 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. The main source of information to apply this method to Singapore is the Household interview travel survey carried out in 2012 (HITS 2012). To determine where the activities are performed, a dataset with more than one
hundred thousand activity locations of Singapore, 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 evoked places 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 first section presents an overview of the whole method. Then, details of the estimation of the logistic regression models are presented in the second section, while the third one elaborates on the Activity agenda generation and the Set of evoked places formation. Section 6.4 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. The last section presents some conclusions, including some hints for future works on this topic.

6.1 Overview of the flexible activities scheduling method

This work focuses on how activity agendas and evoked places can be extracted from common datasets. Figure 6.1 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 (evoked 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 6.1 represent
6.2 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 the Singaporean 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 6.1. The main objective is to find out which socio-demographic and geographic attributes influence the type of places 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 geo-spatial and travel diary survey 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 evoked places and the Activity agenda estimation. With this activity agenda, and the set of evoked places, optimal activity schedules can be found. The Spatio-temporal network method on the bottom-left corner of the figure, finds maximum-utility activity schedules, as explained in Chapter 4. Although, with the activity agenda and the set of evoked places, realistic non-optimal schedules could also be identified using less expensive methods. In the next section 6.2, the binary logistic models are presented in detail. In addition, the Selection of evoked places and the Activity agenda estimation are elaborated in the following 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.
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: Continuous variable
- Gender: Binary dummy variable
- Car availability: Binary dummy variable
- 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) using the database of activity facilities collected for Singapore and travel times estimated with the MATSim Singapore scenario presented in Chapter 2. The general accessibility of a person to a place type is the maximum accessibility among all his/her primary locations. Thus, if for instance 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 model: $\text{Acc}(p_i, t) = \sum_{p_j \in P(t)} (A(p_j)^\beta) e^{\alpha T(p_i, p_j)}$. 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. For a certain combination of $\alpha$ and $\beta$, one accessibility measurement of the primary activity locations of a HITS respondent can be calculated per type of place. Then, a linear regression was made between these accessibilities and the duration of corresponding flexible activities reported by the respondent. The coefficient of determination of the regression indicates how linear is the relation between these variables. The calibration was made by varying the parameters and selecting the combination with the maximum coefficient of determination. Table 6.1 presents calibrated parameters.
6.2. Participation in flexible activity purposes and awareness of place types

- 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

Table 6.1: Calibrated parameters for accessibility calculation by activity type.

<table>
<thead>
<tr>
<th>Place type</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shop high</td>
<td>$-1.6E-3$</td>
<td>1.4</td>
</tr>
<tr>
<td>Shop low</td>
<td>$-1.5E-3$</td>
<td>1.5</td>
</tr>
<tr>
<td>Eat high</td>
<td>$-1.4E-2$</td>
<td>0.8</td>
</tr>
<tr>
<td>Eat low</td>
<td>$-1.2E-2$</td>
<td>0.2</td>
</tr>
<tr>
<td>Recreation</td>
<td>$-1.0E-4$</td>
<td>$-0.4$</td>
</tr>
<tr>
<td>Other home</td>
<td>$-2.0E-3$</td>
<td>0.2</td>
</tr>
<tr>
<td>Park</td>
<td>$-4.2E-3$</td>
<td>$-0.1$</td>
</tr>
<tr>
<td>Civil</td>
<td>$-1.3E-4$</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Figure 6.2 presents the descriptive analysis of respondents in HITS who perform flexible activities.

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,
Figure 6.2: Descriptive analysis of socio-demographic attributes of respondents performing flexible activities in HITS.
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 6.3.

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 following rules were used:

1. When one of the types of the reported trip destination is X, the outcome is 1.
2. When the reported trip purpose (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 6.2 shows some examples of the classification of trips for the "Go to community centre" model.
Figure 6.3: Relation between flexible activities and place types.
Table 6.2: Some examples of the classification of trips for the "Go to community centre" model.

<table>
<thead>
<tr>
<th>Place</th>
<th>Purpose</th>
<th>Age</th>
<th>Gender</th>
<th>...</th>
<th>Out</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Park</td>
<td>rec</td>
<td>40</td>
<td>F</td>
<td>...</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>M</td>
<td>...</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>M</td>
<td>...</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>M</td>
<td>...</td>
<td>1</td>
<td>Social is related with community centre</td>
</tr>
</tbody>
</table>

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" logistic regressions 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 validation. Binary logistic regression requires the modeller to specify the utility functions for each of 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 sections these models are summarized and analysed.

6.2.1 "Go to place of type X" model estimation

As mentioned before the place types modelled were recreational facilities, parks, community centres, homes of others, high-demand shopping places,
Figure 6.4: Some dependencies in HITS of the willingness to travel or not to places of different types 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.

low demand shopping places, high-demand eating places, and low demand eating places. Figure 6.4; 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 is modeled as a weighted sum of the attributes shown in Figure 6.5; For instance, the utility of going to high-demand shopping places is presented in the equation 6.1; the variables included in this function can be found in Figure 6.5;
$U_{\text{shopH}} = \beta_{\text{ageL17}}(\text{ageL17}) + \beta_{\text{ageH17}}(\text{ageH17}) + \beta_{\text{fem}}(\text{fem}) + \beta_{\text{carAv}}(\text{car Av}) + \beta_{\text{chi}}(\text{chi}) + \beta_{\text{ind}}(\text{ind}) + \beta_{\text{mly}}(\text{mly}) + \beta_{\text{accLow}}(\text{accLow}) + \beta_{\text{accHigh}}(\text{accHigh}) + \beta_{\text{hSize}}(\text{hSize}) + \beta_{\text{mainR}}(\text{mainR}) + \beta_{\text{prtnR}}(\text{prtnR}) + \beta_{\text{yngR}}(\text{yngR}) + \beta_{\text{inc0}}(\text{inc0}) + \beta_{\text{inc}}(\text{inc}) + \beta_{\text{hIncB7k}}(\text{hIncB7k}) + \beta_{\text{hIncA7k}}(\text{hIncA7k}) + \beta_{\text{timeH}}(\text{timeH}) + \beta_{\text{timeWB7h}}(\text{timeWB7h}) + \beta_{\text{timeWA9h}}(\text{timeWA9h}) + \beta_{\text{noWork}}(\text{noWork}) + \epsilon$

(6.1)

Logistic regression was used to estimate the coefficients. BIOGEME (Bierlaire, 2013) was employed for this calculation. Figure 6.5 presents the estimated coefficients of the "Go to place of type X" models with their significance levels. For the coefficient estimation, dummy variables were created for categorical variables while linear scaling was applied to continuous variables to limit their ranges to the interval $[0, 1]$.

Interesting findings can be seen in this Figure. Age influences almost all the place types, with positive and negative correlations with different types of places according to the age interval. For instance, age is negative correlated to parks below 24 years and then it becomes positive above this age. Accessibility is one of the strongest predictors, with positive correlations with almost all the place types, except for high accessible low-demand eating places and low accessible high-demand shopping places. Although the first 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 positive 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.
Figure 6.5: Summary of the "Go to place of type X" models estimation. Coefficients in the center and t-test values on the bottom-right.
6.2.2 "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 6.6 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. 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]. The utility of the "Perform eating" models is a linear sum of the attributes shown in Figure 6.7 as shown in Equation 6.2. In this Figure, the optimal coefficient and the corresponding t-test value are presented for each of the 5 models, and each model is displayed in a column.

\[ U_{eat} = \beta_{ageL40}(ageL40) + \beta_{ageH40}(ageH40) + \beta_{fem}(fem) + \]
\[ \beta_{carAv}(carAv) + \beta_{chi}(chi) + \beta_{ind}(ind) + \beta_{mly}(mly) + \]
\[ \beta_{accEatHigh}(accEatHig) + \beta_{accEatLow}(accEatLow) + \]
\[ \beta_{accShopHigh}(accShopHigh) + \beta_{accShowLow}(accShopLow) + \]
\[ \beta_{accHomeO}(accHomeO) + \beta_{hSize}(hSize) + \beta_{mainR}(mainR) + \]
\[ \beta_{prtnR}(prtnR) + \beta_{yngr}(yngr) + \beta_{inc0}(inc0) + \beta_{inc}(inc) + \]
\[ \beta_{hIncB2k}(hIncB2k) + \beta_{hIncA2k}(hIncA2k) + \beta_{timeH}(timeH) + \]
\[ \beta_{timeW}(timeW) + \beta_{noWork}(noWork) + \epsilon \]

(6.2)

Similar to "Go to place of type X" models, interesting findings can be drawn from the results summarized in Figure 6.7. 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 6.3) were included as possible predictors. It’s interesting to notice that the accessibility to high-demand shopping places and community centres is highly and negatively correlated with performing errands. Having a car resulted in negative correlation with
Figure 6.6: Some dependencies in HITS of the performance or non-performance of flexible activities 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.

shopping. Finally, an analysis of the dependencies of the household roles in shopping and social activities was made. As expected, results show that having the younger role is negatively correlated with running errands and shopping. However, it is not obvious that people with this role are negatively correlated with social activities. This could be explained by the strict education culture in Singapore. It is also not expected that having the partner role is negatively correlated with shopping.
Figure 6.7: Summary of the "Perform activity X" models estimation. Coefficients in the center and t-test values on the bottom-right.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Eat</th>
<th>Shop</th>
<th>Social</th>
<th>Recreation</th>
<th>Errands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersect</td>
<td>-3.84</td>
<td>-9.79</td>
<td>-5.97</td>
<td>-10.1</td>
<td>4.67</td>
</tr>
<tr>
<td>Age &lt; 39</td>
<td>3.04</td>
<td>6.95</td>
<td>20</td>
<td>22.39</td>
<td>2.28</td>
</tr>
<tr>
<td>Age &gt; 39</td>
<td>1.92</td>
<td>0.445</td>
<td>0.96</td>
<td>0.284</td>
<td>0.475</td>
</tr>
<tr>
<td>Car availability</td>
<td>0.284</td>
<td>0.113</td>
<td>0.099</td>
<td>0.099</td>
<td>0.74</td>
</tr>
<tr>
<td>Age &lt; 39</td>
<td>0.138</td>
<td>-0.09</td>
<td>-0.6</td>
<td>-0.184</td>
<td>1.42</td>
</tr>
<tr>
<td>Age &gt; 39</td>
<td>-0.138</td>
<td>-0.09</td>
<td>-0.6</td>
<td>-0.184</td>
<td>1.42</td>
</tr>
<tr>
<td>Female</td>
<td>-0.087</td>
<td>-1.51</td>
<td>0.822</td>
<td>0.043</td>
<td>0.043</td>
</tr>
<tr>
<td>Household size</td>
<td>1.04</td>
<td>5.5</td>
<td>0.76</td>
<td>5.5</td>
<td>0.445</td>
</tr>
<tr>
<td>Main role</td>
<td>0.275</td>
<td>0.261</td>
<td>-0.104</td>
<td>2.17</td>
<td>8.14</td>
</tr>
<tr>
<td>Partner role</td>
<td>0.1</td>
<td>1.14</td>
<td>-0.155</td>
<td>-2.08</td>
<td>-0.365</td>
</tr>
<tr>
<td>Younger role</td>
<td>0.06</td>
<td>0.48</td>
<td>-0.269</td>
<td>2.37</td>
<td>-1.26</td>
</tr>
<tr>
<td>Home time</td>
<td>-2.29</td>
<td>2.17</td>
<td>8.14</td>
<td>1.6</td>
<td>-1.88</td>
</tr>
<tr>
<td>Work time</td>
<td>-4.5</td>
<td>-2.33</td>
<td>-4.07</td>
<td>-4.4</td>
<td>-2.22</td>
</tr>
<tr>
<td>No work</td>
<td>-0.743</td>
<td>-4.83</td>
<td>1.08</td>
<td>5.52</td>
<td>0.732</td>
</tr>
<tr>
<td>Income</td>
<td>-0.256</td>
<td>-2.18</td>
<td>-1.65</td>
<td>-1.98</td>
<td>0.01</td>
</tr>
<tr>
<td>No Income</td>
<td>0.046</td>
<td>0.48</td>
<td>0.043</td>
<td>0.46</td>
<td>0.682</td>
</tr>
<tr>
<td>Household income</td>
<td>-1.79</td>
<td>-3.79</td>
<td>-1.11</td>
<td>-2.32</td>
<td>0.594</td>
</tr>
<tr>
<td>Income &gt; 2500</td>
<td>-0.09</td>
<td>-0.42</td>
<td>0.367</td>
<td>4.27</td>
<td>0.709</td>
</tr>
</tbody>
</table>

6.2. Participation in flexible activity purposes and awareness of place types
6.3 Evoked places selection and Activity agenda estimation

In this section, three of the processes mentioned in Figure 6.1 will be explained in detail. The *Distance-based selection of evoked 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. These two processes have to be executed once for each decision maker. Finally, on the bottom of figure 6.1 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).

6.3.1 Evoked places set construction

This procedure returns a fixed number $N$ of evoked places for a decision maker, based on the previously defined socio-demographic attributes. It also returns travel times between each pair of evoked places. The *selection of evoked 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 work proposes two steps for the *selection of evoked places*. The first selection depends only on the types of the places, selecting $F \ast N$ places ($F$ is a factor greater than 1). The second step depends on the geographical dimension and the selection returns $N$ out of $F \ast N$ places. Next the two steps are explained in detail.

**Step 1: Selection by type**  With the utility functions estimated with the "Go to place of type" models, described in the previous section 6.2, 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 (6.3):

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

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

(6.3)

$$\epsilon_t \sim \text{Logistic}$$

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

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 (6.4):

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

(6.4)

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 $F \ast N$ destinations using these values.
Figure 6.8: Distribution of the travel time to destinations ratio from HITS 2012.

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.

**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, people tend to travel to locations near their primary activity locations. Figure 6.8 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 evoked 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.
6.3. Evoked places selection and Activity agenda estimation

Figure 6.9: Travel time kernel density plots of each place type by car mode extracted from HITS 2012.

The extracted travel time density plots for each place type are presented in figure 6.9.

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 $f_{t1}, f_{t2}, \ldots$ will be the final measurement.

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

As mentioned previously, 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 \( a \) is a duration \( d \) and a frequency \( f \). Figure 6.10 shows the distributions of activity duration by type, obtained from HITS 2012. The probability that a person \( y \) performs a flexible activity \( a \) is given by the equation 6.5:

\[
U_a = \beta_{a0}(v_{a0}) + \beta_{a1}(v_{a1}) + \ldots + \beta_{am_a}(v_{am_a}) + \epsilon_a
\]

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

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

\[
P(y_a = 1) = \frac{1}{1 + e^{-(\beta_{a0}(v_{a0})+\beta_{a1}(v_{a1})+\ldots+\beta_{am_a}(v_{am_a}))}}
\]

In this equation \( U_a \) is the utility of performing the activity \( a \), \( \beta_{ai} \) and \( v_{ai} \).
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 previous section 6.2, 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 6.6:

$$P(y_a = 1|y_f = 1) = \frac{P(y_a = 1 \cap y_f = 1)}{P(y_f = 1)} = \frac{P(y_a = 1)}{P(y_f = 1)}$$  (6.6)

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.

### 6.4 Validation experiments and Analysis

The main objective of the experiments carried out is to demonstrate the importance of personal characteristics when scheduling flexible activities. Hence, the following results 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. Schedules for 566 persons were constructed using the algorithm described above. These predictions took 31 seconds using a single-thread java program running on a Windows machine with a common processor (intel i7); the full population of Singapore would takes 8.6 hours at this rate. Then,
for each person in the test set, different maximum utility activity schedules were calculated, but using different selection methods for evoked places and randomly constructed agendas. Figure 6.11 presents an example of the location of the evoked places of a user using four different selection methods.

Figure 6.11: Example of selected evoked places of a decision maker according to his/her primary locations using four selection methods.

Figure 6.12 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, while activity durations made with random information resulted very short. 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 6.13 presents a crucial issue in transportation studies, travel times.
This figure shows comparisons of three travel time distributions for each flexible activity. Blue curves show the travel time distributions of reported trips which start in a main facility and end in a flexible activity location (HITS distributions). Green curves show the distribution of the travel time from the same main locations to all the evoked places of the person where the corresponding activity can be performed (Evoked places distributions).

Figure 6.13: Predicted travel time distributions for the five flexible activities.

Finally, red curves represent distributions of the travel time of trips.
which start from the same main locations, but end in a predicted location from the set of evoked places (Predicted distribution). For almost every flexible activity (except social) the Evoked places distribution contains longer trips than the HITS distribution and the Predicted distribution. This can be explained arguing that evoked places are selected using place type travel time distribution (presented in Figure 6.9) and a certain place type affect several activity types. Furthermore, some evoked places are selected according to different main locations. It is interesting that when optimizing flexible activity chains the selected evoked places produce a more accurate travel time distribution. This is appreciable for shopping, eating and errands activities in this sample.

To assess the correlation of socio-demographic characteristics on the predicted schedules, Figure 6.14 compares the random input prediction with the systematic prediction at 10 different age intervals.

Figure 6.14: 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.

The systematic method is very accurate judging by the high correlation of age in performing recreational activities and going to community centres.
Figure 6.15: Zipf probability mass function of activity chains in HITS and predicted ones.

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.

Finally, the frequency of predicted activity chains can be compared with the reported ones. The chains can be ranked from the most frequent (home-work-home) to the least frequent ones. As shown in Equation 6.7, the Zipf’s law (Newman, 2005) states that this frequency decreases exponentially when the rank grows (Power law). When the logarithm of the frequency is plotted against the logarithm of the rank, these relations are linear with high coefficients of determination. Although this law is empirical and it was found in linguistics, an analysis of activity chain frequencies can be performed.

Figure 6.15 shows that activity chain frequencies decrease as described by the Zipf’s law, both in HITS and in predicted chains. The calculated slope
indicates the exponent characterizing the distribution as shown in Equation 6.7. The two exponents being greater than 1 indicates that the distributions are very heavy-tailed, i.e. that few activity chains are very frequent and most of the chains occur one or two times. The HITS distribution exponent (1.95) being larger than the predicted distribution exponent (1.71) means that the most popular reported activity chains are more frequent than the most popular predicted activity chains, and the least popular reported activity chains are less frequent than the less popular predicted activity chains. This shows that the method schedules more combinations of flexible activities than the reported ones.

\[
f(r; s, N) = \frac{1/r^s}{\sum_{n=1}^{N} 1/n^s} \\
f(r; s, N) = \frac{1}{\sum_{n=1}^{N} 1/n^s} \frac{1}{r^s} \\
f(r) = \frac{k}{r^s} \\
\log(f(r)) = \log(k/r^s) \\
\log(f(r)) = \log(k) - s(\log(r))
\]

Where \( f \) is the relative frequency of an element, \( r \) is the rank of the element, \( N \) is the number of different elements, and \( s \) is the exponent characterizing the distribution.

### 6.5 Conclusions and discussion

In the same way as the previous chapter, the main goal of this chapter about scheduling personalized flexible activity chains, was also to design models based on available data from Singapore. Flexible activity patterns of this city where 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 schedul-
ing 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.

Observed travel time distributions in HITS were compared with predicted travel time distributions for each flexible activity. It was found that, although travel time distributions generated using all evoked places contain longer travel times, travel time distributions generated using the scheduled evoked places are more accurate. This could model the idea that, although there are places in people’s choice set which are far from the current location, the selected places are closer.

When analyzing reported and generated activity chains, it was found that both HITS chains and predicted chains present very heavy-tailed distribution as expected by the Zipf’s law. However, the difference in the exponents which characterize these two distributions indicates that the method schedules more combinations of flexible activities than the reported ones.

To improve prediction capabilities, more research must be carried out on the activity utility function of the algorithm. Currently it depends in two parameters only (typical duration and frequency), and observed distributions could be included. 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.
Chapter 7

Weekly simulations in Singapore

In this chapter, the methods and models developed and explained in Chapter 5 and Chapter 6 are used to set-up an activity-based weekly demand of Singapore. Then, the activity-based weekly demand is simulated, analyzed and validated against the Household Interview Travel Survey perform in 2012 (HITS). Three key elements are used for this simulation: (i) an agent-based simulation platform (MATSim), (ii) its extension described in Chapter 3, and (iii) the multi-activity scheduling algorithm presented in Chapter 4. Although the models are estimated and validated with the same data (HITS), the objective of these experiments is to show that the proposed abstraction works, and future scenarios can be simulated as a result of these new techniques.

This chapter starts with an overview of the activity demand generation process, followed by two sections describing in detail the two main pillars of the proposed approach: **Fixed activities weekly skeletons** and **Flexible activity agendas, and evoked places**. A fourth section describes the simulation results and performance, followed by analysis of the results comparing them with HITS. Finally, a section for general conclusions closes the chapter.

### 7.1 Overview

As explained in Chapter 2, the activity-based demand for agent-based simulations consists of a fully detailed activity chain for each agent in a population during a particular period. Thus, to generate an activity-based
demand, it is necessary to specify the type of activities of the chain, their sequence, their start time and duration, and the location where each activity is performed for each agent. Assuming there are only two activities per day, a weekly activity demand for one agent would consist of a sequence of 21 activities. Assuming a very simple activity model with only four types of activities (i.e. home, work, study and other), the number of possible activity chains adds up to $268,6$. By increasing the number of activities per day from 2 to 3, the number of combinations grows to $4.4E12$. Therefore, to assign realistic weekly activity chains to thousands or millions of agents, i.e. to generate a weekly activity-based demand, is a computationally intractable problem.

This work proposes a new definition of **weekly activity-based demand** as mentioned in Chapter 2. First, as introduced in Chapter 5, activities are categorized into **Fixed** and **Flexible** activities according to how mandatory it is to perform the activity at a certain time and place. Thus, for each agent, the new activity-based demand is composed of two parts: (i) **Fixed activities weekly skeletons**, and (ii) **Flexible activity agendas and evoked places**.

A **Weekly skeleton** is a sequence of **Fixed activities** performed during one week. To assign weekly skeletons, agents are divided into three types: workers, students and others. The others refer to people who do not carry out any main activity. As a worker performs up to two fixed activities (home, work) per day, the total number of possible **weekly skeletons** is merely $16,384$. For students, the number of combinations is the same, while there is only one possible combination for the others. As a result of these simplifications, the number of combinations is reduced by sixteen thousand times against the standard approach.

When a **fixed activity skeleton** has defined activity times, free time windows are revealed between consecutive fixed activities. **Flexible activity agendas** and **evoked places** are used to schedule on-the-fly flexible activities during these windows. Although the time and location of flexible activities are not specified in advance, these activities will be scheduled during the simulation with updated information. This approach is also useful to include several flexible activity types in the activity-based model. Although the number of activities is limited by the complexity of the scheduler, the tractability of the scheduling problem is no longer dependent on the total period of the simulation, but on the maximum size of a free time window.
The beginning of this chapter explains how to assign this new activity-based demand definition to the synthetic population generated for the Singapore scenario introduced in Chapter 2. The synthetic population generation process is not explained in this thesis but details can be found in Erath et al. (2012). The information for each agent in the synthetic population is:

- Sociodemographics: Age, Gender, Occupation, Household size, Income, etc.
- Home location
- Main activity location: If the agent is a worker or a student the location where it performs its main activity is expected to be known.
- Car availability

This scenario and the new models are based on data which is commonly available in many cities:

1. Household interview travel survey (HITS 2012): Revealed trips of approximately 1% of the Singaporean population.
2. Public transport smart card data (CEPAS): Transactions of 97% of public transport users.
3. Singapore Statistics (SingStat): Aggregated census data by geographic zones at different scales.

In addition, the scenario includes a database of activity facilities of the city. Several sources of information were used to generate this database and details can be found in Erath et al. (2012).

Figure 7.1 shows the processes designed to generate this weekly activity demand. Black boxes represent the processes mentioned above, which are assumed to be previously executed. Red shapes display processes to produce final outcomes; these are more specifically XML files that can be used to run a MATSim instance. Colored borders highlight processes that use directly one of the mentioned Singaporean datasets. Boxes with colored backgrounds are processes explained in previous chapters of this thesis. Finally, the number at each box represents one of the possible execution orders of the weekly activity demand generation.

Next sections will describe these processes, and will present the results validating them against the original data.
7.2 Fixed activities weekly skeletons

The proposed weekly demand generation starts with the main activity weekly patterns recognition which is explained in detail in Chapter 5. In these processes (Boxes 1, 2 and 3 in Figure 7.1), HITS is used to estimate activity type choice models, assign types to detected activities in CEPAS data, and find weekly patterns of frequent public transport users for managing a small number of work and study activity behaviors.

Next, in order to assign one of the clustered patterns to each agent of the synthetic population, six new processes are designed and implemented. These are represented by boxes 4, 5, 6, 7, 8 and 9 in Figure 7.1. Two important assumptions have to be made to apply these models:

- Fixed activity weekly patterns of frequent public transport users in Singapore are the same of the rest of the population.
- People living in the same planning area have similar fixed activity
weekly patterns.

Although there are many arguments against these two assumptions, it is necessary to adopt them in order to use the results found in Chapter 5 to model fixed activities of the whole synthetic population. The first assumption is mandatory because weekly public transport data is the only source of continuous multi-day data available for this work. The second assumption is also a must because the only sociodemographic information of public transport users in the CEPAS system which can be inferred is based on the census zone where they live.

The six processes of the weekly skeleton assignment can be divided in two: (i) the development of the weekly skeleton models represented by boxes 4, 5 and 6, and (ii) the application of those models to the synthetic population represented by boxes 7, 8 and 9 in Figure 7.1. Following, these two parts are described and the results are validated against the original data.

### 7.2.1 Weekly skeleton models

Sociodemographic characteristics of people determine fixed schedules of their mandatory activities. Among them, occupation is the most representative feature. For certain occupations, workers are required to stay more time at their fixed working places. The size of the household could also influence the time a person intends to be at home. In this subsection, the significance of sociodemographic characteristics on weekly activity patterns is studied and modeled.

Based on sociodemographic data from the census, *Weekly skeleton models* must be estimated to categorize people in one of the most popular clusters of fixed activity patterns found in Chapter 5. For this estimation, frequent public transport users that belong to the most popular clusters are used as a training dataset. The estimation has three steps:

1. Find the planning area where the home location of users in the training dataset is located.
2. Assign each user in the training dataset the socio-demographic proportions of his/her corresponding planning area as fuzzy attributes.
3. Feed the corresponding *Weekly skeleton model* with observations composed by fuzzy attributes as input and corresponding cluster as output.
Thus, the value of each attribute of users in the training dataset is a set of probabilities, i.e. the ethnicity of the user with id="A10" is 30% Chinese, 45% Indian, 20% Malay and 5% Other. Discretization of continuous variables is necessary. Similar to the procedures presented in Chapter 5, two models are estimated, one for workers and the other for students.

The random forest technique (Ho, 1995) is employed to estimate the influence of these attributes in the weekly pattern of workers and students. To train this machine learning algorithm, a controlled number of decision trees are created with different samples of the training dataset. When the random forest model is trained, it can be used to predict the output of new observations. Given the input data of a new observation, each decision tree returns an independent result. This allows this algorithm to return fuzzy classifications.

As some of the clusters mentioned above include only a small number of observations, the oversampling technique is used to generate random new observations for these clusters. Existing observations with the same output are used for this generation (Han et al., 2005). When the size of the clusters becomes similar, the performance of the random forest algorithm improves.

### 7.2.1.1 Workers weekly skeleton model

The attributes used for the Workers weekly skeleton model are:

- **Age**: From 0 to 65 years, there are 13 intervals of 5 years, and a 14th interval for more than 65 years old.
- **Gender**: Male or Female.
- **Ethnicity**: Chinese, Indian, Malay or Other.
- **Occupation**: Ten types: Agricultural and Fishery, Technicians, Cleaners and laborers, Clerical workers, Machine operators, Production craftsmen, Professionals, Managers, Sales workers or Others.
- **Car availability**: Yes or No

For the workers model, a sample of 5252 observations is obtained from users in the most popular 17 clusters of working weekly patterns. A sample of 20% is left for testing and 80% of the observations are employed to train the random forest. There are approximately 1800 observations in the biggest cluster of the training set. Every other cluster is over-sampled to obtain a similar number of observations. A total number of 21,390 observations are finally used to train the random forest.
The main predictors of the model are Ethnicity and Age. This is measured by the mean decrease in Gini index. The Gini index measures how unequal is a characteristic in a group. When the Gini index is 0 the characteristic (cluster) is completely equal and when the Gini is 1 the characteristic is unequal. In this description the Gini index will be expressed in the $[0–1000]$ interval for simplicity. To select which is the best rule and predictor to apply to a group while a decision tree is generated, the difference between the Gini index in the original group and the average of the Gini indices in the resulting subgroups is calculated. If the predictor is good that difference (decrease Gini) is big, because the subgroups are more equal internally. To measure the mean decrease in Gini of a predictor, the average of the decrease in Gini of that predictor in all the trees is calculated. The Indian race predictor decreases Gini by 499 units on average. The 15 to 25 years old predictor reduces the Gini by 484 units. The most influential occupation is agricultural and fishery, making the Gini factor 280 units lower.

Figure 7.2 shows the actual frequency of workers by cluster in the testing dataset in comparison with the frequency predicted by the model. Although the Out-Of-Bag error is 56%, the predicted cluster distribution of workers in the testing set has an error of 17.5% only. The Out-Of-Bag error is calculated by measuring the wrong predictions each time a tree is generated, with observations which were not used to train that tree. This error commonly higher because one tree doesn’t have the same prediction capability than the whole forest. The model has a decent prediction capability, taking into account that these predictions were made with only 5 sociodemographic characteristics and the values of these attributes are distributions (fuzzy attributes). It also generates a similar cluster distribution as observed in Figure 7.2.

### 7.2.1.2 Students weekly skeleton model

For the students model, a sample of 1663 observations are obtained from users in the most popular 17 clusters of studying weekly patterns. A sample of 20% is left for testing and 80% of the observations are used to train the random forest. There are approximately 300 observations in the biggest cluster of the training set. Every other cluster is over-sampled to obtain a similar number of observations. A total number of 4049 observations are finally obtained to train the random forest.
Figure 7.2: Real vs. estimated distribution of workers in the testing data by the most popular clusters.

The main predictors of the model were also Ethnicity and Age. The Indian ethnicity predictor decreases the Gini index in 112 units, while being studying between 20 and 30 years decrease it 110 units in average. Figure 7.3 shows the frequency of students by cluster in the testing dataset compared with the frequency predicted by the model. The Out-Of-Bag error is 68%, and the predicted cluster distribution of students in the testing set has an error of 39.8%. As explained before, these high errors are obtained because the number of predictors is limited and many student clusters are quite similar. However, the frequency of these clusters is well predicted according to Figure 7.3.

7.2.2 Weekly skeleton assignment

To schedule fixed activities for the agents in the synthetic population introduced in Chapter 2, the idea is to categorize each agent in one of the
presented clusters, and obtain start time and duration of these activities from the cluster times. Thus, for each cluster \(c\), a joint empirical distribution of start time and duration of the home activity \(H^d_c\) is estimated for each day of the week \(d\). For the 17 work clusters, joint empirical distributions of start time and duration of the work activity \(W^d_c\) are estimated for the days of the week when a working activity is included. For example, for a cluster with 5 working days, 7 joint distributions are estimated for home activities, and 7 joint distributions are estimated for working activities. In the same manner, for the 17 study clusters, joint distributions are estimated for the start time and duration of study activities \(S^d_c\).

The random forests estimated in the previous subsection, are now used to classify workers and students of the synthetic population in one of the clusters presented. Given a working agent \(a\) assigned to a cluster \(c_i\), random start time and duration from \(H^d_{c_i}\) are sampled to define its home activity for the day \(d_j\). Then, random start time and duration from \(W^d_{c_i}\) are sampled to define the working time for agent \(a\). Working activities have priority when there is an intersection in time. Thereby, home and work activities can

---

Figure 7.3: Real vs. estimated distribution of students in the testing data by the most popular clusters.
be scheduled for all working agents, for every day in the week, and time windows of free time are defined between fixed activities. Similarly home and study activities can be scheduled for all studying agents using $H_c^d$ and $S_c^d$, and prioritizing the study activity when there is an intersection in time. For agents without a main activity, home activities are scheduled using $H_c^d$.

It is important to highlight that the main source of these activity times are detected home, working and studying activities in the CEPAS data. These activities were detected using reported activities reached by public transport from HITS as explained in Chapter 5. According to the assumption mentioned above: "Fixed activity weekly patterns of frequent public transport users in Singapore are the same than weekly patterns of the rest of the population", it is expected to obtain activity times for the synthetic population which are similar to times reported in HITS by respondents traveling by any mode.

Figure 7.4 and figure 7.5 show the resulting distributions of start time and duration of home activities by day of the week, compared with the same distributions reported in HITS (HITS). Although the mean of the start time distributions are similar, there are more reported home activities starting in the middle of the day. This can be explained arguing that the current model just includes one home activity per day, and people report going back home in the middle of the day in HITS. This also explains the short reported home activities in figure 7.5, which are not predicted by this model. Home activities over 18 hours long are also underestimated. This must be due to frequent public transport users who do not stay that long at home. For weekends, home activities are estimated longer than the reported, especially on Saturdays.

Similarly, figure 7.6 and figure 7.7 show the resulting distributions (Simulated) of start time and duration of working activities by day of the week, compared with the same distributions reported in HITS (HITS). Estimated distributions are very accurate, with some more short work activities reported in HITS. This can be due to the activity detection model which categorizes short work activities as non-work.

Finally, figure 7.8 and figure 7.9 show the same distribution comparisons for studying activities. Start times are accurate with some more activities reported to start after 10 a.m. However, the duration of study activities is systematically overestimated. The reason must be related to extra-curricular activities of students. They report shorter study activities because they
Figure 7.4: HITS vs. estimated start time distribution of the home activity.

Figure 7.5: HITS vs. estimated duration distribution of the home activity.
Chapter 7. Weekly simulations in Singapore

Figure 7.6: HITS vs. estimated start time distribution of the work activity.

Figure 7.7: HITS vs. estimated duration distribution of the work activity.
perform other activities at the same educational facilities. But in the public transport transactions, study activities appear longer covering formal studying and extra-curricular activities.
7.3 Flexible activity agendas and evoked places

This section explains how the models presented in Chapter 6 are used to assign an Activity Agenda and a Set of evoked places to agents in the synthetic population of Singapore. These models are based in a relation between flexible activity types and location types.

Given the activity locations database collected for Singapore and introduced in Chapter 2, Table 7.1 show the proposed location classification. In Singapore, postal codes are assigned to (just about) every building or facility of the city. In the locations database, several destinations can be assigned to the same postal code or facility. Hence, the same facility can be classified in several types, and the number of destinations for each type defines an attraction index for the facility. For example a shopping mall should be modeled as a facility with a high attraction index for the types SHOP and EAT, and a moderate attraction index for the types NEED and BUSINESS. In Figure 7.1, boxes 12 and 13 represent the processes of classifying facilities by type and writing the MATSim facilities file, which
Table 7.1: Classification of locations according to collected points of interest in Singapore.

<table>
<thead>
<tr>
<th>Type</th>
<th>Points of interest</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHOP HIGH</td>
<td>Shopping mall, Super market, Department store</td>
<td>Locations for shopping</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with high demand.</td>
</tr>
<tr>
<td>SHOP LOW</td>
<td>Stores, Mini market, Florist, Car dealer, Bakery, Art gallery</td>
<td>Locations for shopping</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with low demand.</td>
</tr>
<tr>
<td>EAT HIGH</td>
<td>Shopping mall, Food court, Hawker center</td>
<td>Locations for eating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with high demand.</td>
</tr>
<tr>
<td>EAT LOW</td>
<td>Cafe, Restaurant, Bakery, Meal take away, Fast food</td>
<td>Locations for eating</td>
</tr>
<tr>
<td></td>
<td></td>
<td>with high demand.</td>
</tr>
<tr>
<td>BUSINESS</td>
<td>Bank, Court, Agency, Embassy, Casino, ATM</td>
<td>Locations for performing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>activities related with money.</td>
</tr>
<tr>
<td>NEED</td>
<td>Fire, police and gas stations, Contractors, Storage, Car rental, Post office, Hair care, Library</td>
<td>Location for personal services.</td>
</tr>
<tr>
<td>CULTURAL</td>
<td>Zoo, Aquarium, Museum, Cinema, Art gallery, Library</td>
<td>Locations for arts and</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cultural activities.</td>
</tr>
<tr>
<td>FUN</td>
<td>Bar, Casino, Night club, Bowling alley, Amusement park</td>
<td>Locations for social</td>
</tr>
<tr>
<td></td>
<td></td>
<td>entertainment.</td>
</tr>
<tr>
<td>SPORT</td>
<td>Park, Spa, Bowling alley, Stadium, Natural, Gym</td>
<td>Locations for practicing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sports</td>
</tr>
<tr>
<td>HEALTH</td>
<td>Hospital, Clinic, Dentist, Physiotherapist, Practitioner</td>
<td>Locations for health services</td>
</tr>
<tr>
<td>RELIGION</td>
<td>Church, Mosque, Temple, Synagogue, Cemetery</td>
<td>Places of worship</td>
</tr>
</tbody>
</table>

is used as input information for the simulation model.

According to secondary activities defined for the Singapore scenario, and the locations classification in Table 7.1, the proposed relation between flexible activities and location types is presented in Figure 7.10.

As explained in Chapter 6, the next step consists of estimating models which predicts which flexible activities are performed by a person, and which location types a person travels to. As shown in boxes 16, 17 and 18 of Figure 7.1, these models use the household interview travel survey (HITS)...
as input data. As accessibility measurements are one of the best predictors, the box 14 represents the process of calculating accessibility indices for the main locations of the agents in the scenario. For each main location, accessibility by car and public transport are calculated for each location type. In Ordóñez Medina (2016a) different accessibility measurements are tested for this purpose.
Then, these flexible activities models are applied to the agents in the synthetic population of Singapore to obtain individual activity type probabilities. These probabilities are used to define *Activity agendas* for each agent. These two processes are represented by boxes 19 and 20 of Figure 7.1. Next, place type models are used to estimate place types probabilities, which are necessary to define a *Set of evoked places* for each agent. This is represented by boxes 21 and 22. Finally, the box 23 represents a process to write the *Activity agenda* probabilities and the *Sets of evoked places* in a MATSim person attributes file.

Next subsections describe the estimation of the mentioned predictive models, and present validations using a test sample which was not used for estimation. In section 7.5, the accuracy of the *Activity agendas* and the *Sets of evoked places* is demonstrated when these are used to schedule flexible activities on-the-fly during the MATSim weekly simulation.

### 7.3.1 Activity type and Location type models

In Chapter 6, logistic regressions are proposed to estimate activity type and location type probabilities given sociodemographic characteristics. These models are trained using observations in HITS. Now, for the Singapore scenario, the random forest approach is used to estimate the same output. As the relationships between sociodemographic characteristics and the desired probabilities are rarely linear, and many of these characteristics are nominal variables, it is expected that the random forest approach produces better results. The decision trees of a random forest are designed for nominal variables, and continuous variables can be split into desired intervals. The sociodemographic characteristics used for these models are the same mentioned in Section 6.2. Several accessibility definitions were evaluated as predictors. Table 7.2 displays the six proposed accessibility definitions. In Ordoñez Medina (2016a), comparisons of their prediction capabilities are presented. Following, results of the random forests’ estimation and a validation against a test dataset from HITS are described.

#### 7.3.1.1 Activity type random forest models

As shown in Figure 7.10, six flexible activities are included in this simulation model. For each activity type two binary random forest model are estimated. The first models include observations in HITS when no main activity is
Chapter 7. Weekly simulations in Singapore

Table 7.2: Accessibility definitions tested for activity type and location type models.

<table>
<thead>
<tr>
<th>Id</th>
<th>Description</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Travel time to all type-locations weighted by size of the location</td>
<td>( A(i, t) = \sum_{j \in P(t)} t_{ij} \cdot O_j \sum_{j \in P(t)} O_j )</td>
</tr>
<tr>
<td>B</td>
<td>Attractiveness of all type-locations weighted by inverse travel time</td>
<td>( A(i, t) = \frac{\sum_{j \in P(t)} O_j \cdot \frac{1}{t_{ij}}}{\sum_{j \in P(t)} \frac{1}{t_{ij}}} )</td>
</tr>
<tr>
<td>C</td>
<td>Attractiveness of 50 best type-locations weighted by inverse travel time</td>
<td>( A(i, t) = \frac{\sum_{j \in P(t)} 50 O_j \cdot \frac{1}{t_{ij}}}{\sum_{j \in P(t)} \frac{1}{t_{ij}}} )</td>
</tr>
<tr>
<td>D</td>
<td>Attractiveness of all type-locations weighted by tt distribution</td>
<td>( A(i, t) = \frac{\sum_{j \in P(t)} O_j \cdot D_t(t_{ij})}{\sum_{j \in P(t)} D_t(t_{ij})} )</td>
</tr>
<tr>
<td>E</td>
<td>Attractiveness of 50 best type-locations weighted by tt distribution</td>
<td>( A(i, t) = \frac{\sum_{j \in P(t)} 50 O_j \cdot D_t(t_{ij})}{\sum_{j \in P(t)} D_t(t_{ij})} )</td>
</tr>
<tr>
<td>F</td>
<td>Attractiveness of 5 best type-locations weighted by tt distribution</td>
<td>( A(i, t) = \frac{\sum_{j \in P(t)} 5 O_j \cdot D_t(t_{ij})}{\sum_{j \in P(t)} D_t(t_{ij})} )</td>
</tr>
</tbody>
</table>

\( A(i, t) \) is the accessibility of place \( i \) to places of type \( t \). \( t_{ij} \) is the travel time from place \( i \) to place \( j \). \( O_j \) is the attractiveness of place \( j \). \( P(t) \) is the set of places of type \( t \). \( D_t(t_{ij}) \) is the probability density function of travel time constructed with HITS trips to places of type \( t \).

performed by the respondent during the rebuilt day; this normally happens on weekends. The second models are estimated with observations in HITS when the respondent performs main fixed activities: work or study. Given a flexible activity \( a \), each observations in HITS is classified as positive if the respondent performs the activity \( a \) or negative if he/she does not. As there are much more negative than positive observations, oversampling techniques were used to train the random forest models. Table 7.3 describes the performance of the twelve estimated classification models. For each case 20% of the observation were left for validation.

The accessibility definition with best performance is a weighted sum of the attractiveness of all locations of a certain type, where the weight is the inverse of the travel time. It corresponds to the \( B \) identifier in Table 7.2. For some models such as Errands and Religion, a local accessibility resulted a better predictor (identifier \( F \)). It is interesting that accessibility indices result in better predictors for models with observations including a main activity. For example, comparing the two Medical activity models,
Table 7.3: Performance of activity type random forest models.

<table>
<thead>
<tr>
<th>Activity</th>
<th>With main activity</th>
<th>Accessibility index</th>
<th>Training data error</th>
<th>Testing data error</th>
<th>Best predictors (Decrease Gini)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shop</td>
<td>No</td>
<td>B</td>
<td>11.52%</td>
<td>31.57%</td>
<td>Home time (178) Main income (108) Acc ShopL car (82)</td>
</tr>
<tr>
<td>Shop</td>
<td>Yes</td>
<td>A</td>
<td>9.93%</td>
<td>28.90%</td>
<td>Acc Fun car (79) Acc Business car (58) Income (57)</td>
</tr>
<tr>
<td>Eat</td>
<td>No</td>
<td>B</td>
<td>10.42%</td>
<td>27.95%</td>
<td>Main income (181) Home time (128) Acc EatL car (120)</td>
</tr>
<tr>
<td>Eat</td>
<td>Yes</td>
<td>A</td>
<td>9.82%</td>
<td>39.61%</td>
<td>Acc EatH car (39) Acc Business car (39) Acc Cultural car (39)</td>
</tr>
<tr>
<td>Errands</td>
<td>No</td>
<td>B</td>
<td>3.69%</td>
<td>15.44%</td>
<td>Home time (470) Age (276) Main income (232)</td>
</tr>
<tr>
<td>Errands</td>
<td>Yes</td>
<td>F</td>
<td>1.64%</td>
<td>13.31%</td>
<td>Work time (93) Main income (62) Acc Need car (57)</td>
</tr>
<tr>
<td>Medical</td>
<td>No</td>
<td>B</td>
<td>2.25%</td>
<td>6.65%</td>
<td>Home time (528) Main income (367) Age (327)</td>
</tr>
<tr>
<td>Medical</td>
<td>Yes</td>
<td>B</td>
<td>0.33%</td>
<td>1.30%</td>
<td>Work time (174) Main income (126) Acc Health car (84)</td>
</tr>
<tr>
<td>Religion</td>
<td>No</td>
<td>B</td>
<td>1.47%</td>
<td>5.41%</td>
<td>Home time (532) Age (382) Indian (257)</td>
</tr>
<tr>
<td>Religion</td>
<td>Yes</td>
<td>F</td>
<td>1.04%</td>
<td>5.52%</td>
<td>Age (85) Acc cultural pt (81) Main income (72)</td>
</tr>
<tr>
<td>Rec</td>
<td>No</td>
<td>B</td>
<td>4.38%</td>
<td>15.33%</td>
<td>Main income (260) Home time (225) Age (194)</td>
</tr>
<tr>
<td>Rec</td>
<td>Yes</td>
<td>B</td>
<td>1.76%</td>
<td>10.06%</td>
<td>Age (138) Acc Cultural pt (85) Main income (67)</td>
</tr>
</tbody>
</table>
when a main activity is included, the accessibility to Health locations is one of the best predictors, but when no main activities are performed this predictor is not significant. A similar result is found comparing the two Recreational activity models when observing the accessibility to Cultural places. As expected Age plays an important role in the Medical, Religion and Recreational activity models, while the Household income is significant in all the models.

These models are applied to each agent of the synthetic population. Thereby, each agent starts the simulation with twelve new numerical person attributes which are used each time the agent schedules flexible attributes. Additionally, observed duration and multi-day start time distributions of each flexible activity are included in each agent Activity agenda. In Section 7.5, simulated schedules are validated in time and space against HITS records.

### 7.3.1.2 Location type random forest models

Location type models are estimated according to the relation displayed in Figure 7.10. As people perform shopping, eating, errands and recreational activities in places of different types, four random forest models are estimated in a similar way than the logistic regressions presented in Chapter 6. As the activity type models presented above, it is expected that the random forest classification method works better than logistic regressions, because linear relations are rarely found and the majority of predictors are nominal.

Records in HITS with trips to the four mentioned flexible activities are the input data for these models. For example for the recreational model, each trip with a recreation purpose in HITS is classified as trip to a Fun place, to a Cultural place or to a Sport place. Eighty percent of these observations are used to train the recreational activity random forest, and twenty percent are left for validation. Table 7.4 presents the summary of the results of the four location type random forest models:

For the location type random forest models, the accessibility definition with better performance is again the weighted sum of the attractiveness of all the locations of a certain type. The errands model is the only one where the local accessibility performs better. The recreational model resulted with an outstanding low error compared with the other models. Table 7.5:
Table 7.4: Performance of activity type random forest models.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Place types</th>
<th>Accessibility index</th>
<th>Training data error</th>
<th>Testing data error</th>
<th>Best predictors (Decrease Gini)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shop</td>
<td>Shop high</td>
<td>B</td>
<td>11.34%</td>
<td>19.52%</td>
<td>Home time (198)  Age (98)  Main income (74)</td>
</tr>
<tr>
<td></td>
<td>Shop low</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eat</td>
<td>Eat high</td>
<td>B</td>
<td>15.8%</td>
<td>25.97%</td>
<td>Home time (57)  Acc EatL pt (47)  Acc EatH car (45)</td>
</tr>
<tr>
<td></td>
<td>Eat low</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Errands</td>
<td>Business</td>
<td>F</td>
<td>14.03%</td>
<td>29.23%</td>
<td>Age (16)  Main income (10)  Acc Need Car (10)</td>
</tr>
<tr>
<td></td>
<td>Need</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rec</td>
<td>Sport</td>
<td>B</td>
<td>1.52%</td>
<td>6.67%</td>
<td>Acc Fun pt (60)  Main income (51)  Home time (39)</td>
</tr>
<tr>
<td></td>
<td>Fun</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.5: Real vs. Predicted matrix of the *recreational* locations model

<table>
<thead>
<tr>
<th>Training set</th>
<th>Pred / Real</th>
<th>Cultural</th>
<th>Fun</th>
<th>Sport</th>
<th>Testing set</th>
<th>Pred / Real</th>
<th>Cultural</th>
<th>Fun</th>
<th>Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pred / Real</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cultural</td>
<td>306</td>
<td>1</td>
<td>5</td>
<td></td>
<td>Cultural</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Fun</td>
<td>2</td>
<td>315</td>
<td>3</td>
<td></td>
<td>Fun</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Sport</td>
<td>0</td>
<td>3</td>
<td>288</td>
<td></td>
<td>Sport</td>
<td>2</td>
<td>0</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

shows the Real vs. Predicted matrix for the training set and the testing set. Even with three possible outcomes, the prediction accuracy of the generated random forest predicts is 93%. The household income is one of the best predictors for all the models. Accessibility is decisive for the *eating* model, and important for the *errands* and *recreational* model. Finally, age is a good predictor for the *shopping* and *errands* location models.

These random forest models estimate the willingness of a person to travel to locations of certain types. When applying these models to the agents of the synthetic population, the resulting probabilities are used to assign a *Set of evoked places* to each agent. The method of selection is described in Section 6.3. As travel times distributions are also used for this
Figure 7.11: Travel time distribution comparison between observed car trips in HITS and selected *Evoked places* by activity type.

Selection, Figure 7.11 and Figure 7.12 show the comparison between the selected known places travel time distributions by car and public transport respectively. As the selected evoked places emulate the same observed travel time distributions, agents will travel maintaining these distributions in the simulation.
7.4 Weekly simulation

7.4.1 Set up

The weekly simulation is executed with a ten percent sample of the Singaporean synthetic population. The flow capacity factor is set to 0.1 and the storage capacity factor is set to 0.3. According to the simulation preparation presented above, Table 7.6 summarizes the initial conditions for this experiment.

As explained in Chapter 3, agents start with incomplete plans the first iteration. In an incomplete plan or skeleton, only fixed activities are specified. This means no information of trips or flexible activities is given. This is decided by every agent by means of the scheduling algorithm explained in Chapter 4 during the first iteration. For the second one, a sample of agents is selected to start the simulation with incomplete plans, and the rest executes the complete plan generated in the first iteration. According to previous experiences with MATSim, 30% is a good proportion of agents to be selected for a plan modification. The same procedure is
Table 7.6: Initial conditions of MATSim weekly simulation.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario</td>
<td>Singapore</td>
</tr>
<tr>
<td>Sample</td>
<td>10%, 371,996 agents</td>
</tr>
<tr>
<td>Simulated time</td>
<td>0:00 - 174:00, 7 days plus 6 hours</td>
</tr>
<tr>
<td>Iterations</td>
<td>100</td>
</tr>
<tr>
<td>Fixed activities</td>
<td>home, 63 work activities, 87 study activities</td>
</tr>
<tr>
<td>Work model</td>
<td>17 clusters with different activities per day of week</td>
</tr>
<tr>
<td>Study model</td>
<td>17 clusters with different activities per day of week</td>
</tr>
<tr>
<td>Flexible activities</td>
<td>shop, eat, errands, rec, medical, religion</td>
</tr>
<tr>
<td>Network modes</td>
<td>car</td>
</tr>
<tr>
<td>Teleported modes</td>
<td>public transport, taxi, company bus, school bus, walk</td>
</tr>
<tr>
<td>Number of facilities</td>
<td>105, 412</td>
</tr>
<tr>
<td>Home facilities</td>
<td>86, 027</td>
</tr>
<tr>
<td>Work facilities</td>
<td>16, 838</td>
</tr>
<tr>
<td>Study facilities</td>
<td>368</td>
</tr>
<tr>
<td>Flexible activities facilities</td>
<td>7, 041</td>
</tr>
<tr>
<td>Replanning Modules</td>
<td>None</td>
</tr>
<tr>
<td>Within-day planning</td>
<td>Flexible activities and trips, utility maximization</td>
</tr>
<tr>
<td>Population with skeleton</td>
<td>100% first iteration, 30% other iterations</td>
</tr>
<tr>
<td>Memory size of agents</td>
<td>1 Plan</td>
</tr>
<tr>
<td>Scoring function</td>
<td>Extension of Charypar-Nagel function</td>
</tr>
<tr>
<td>Total number of threads</td>
<td>20</td>
</tr>
<tr>
<td>Number of threads for <em>MobSim</em></td>
<td>4</td>
</tr>
</tbody>
</table>

Each agent also starts with a set of *Evoked places* and an *Activity agenda*. In this model, each agent knows 5 places to shop, 5 places to eat, 3 places to run errands, 4 places for recreation, 1 location for medical needs and 1 place of worship. *Activity agendas* include 3 elements for each flexible activity to decide if it can be performed or not. This decision is modeled by the function shown in line 23 of Algorithm 2 in Chapter 4. The 3 elements for each activity are: (i) personal performing probability, (ii) start time distribution for each day of the week and (ii) duration distribution for each day of the week.
Table 7.7: Calibrated mode parameters of weekly simulation.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Marginal utility of time</th>
<th>Marginal utility of distance</th>
<th>Constant utility</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public transport</td>
<td>-2.1 utils/hour</td>
<td>-0.5 utils/km</td>
<td>-1.4 utils</td>
<td>Slow and cheap mode</td>
</tr>
<tr>
<td>Taxi</td>
<td>-1.0 utils/hour</td>
<td>-4.0 utils/km</td>
<td>-2.7 utils</td>
<td>Fast and expensive mode</td>
</tr>
<tr>
<td>Company bus</td>
<td>-2.0 utils/hour</td>
<td>-0.0 utils/km</td>
<td>-0.0 utils</td>
<td>Just for trips from and to work</td>
</tr>
<tr>
<td>School bus</td>
<td>-2.0 utils/hour</td>
<td>-0.0 utils/km</td>
<td>-1.3 utils</td>
<td>Just for trips from and to study</td>
</tr>
<tr>
<td>Walk</td>
<td>-11.0 utils/hour</td>
<td>-0.0 utils/km</td>
<td>-0.0 utils</td>
<td>Very slow and free mode</td>
</tr>
</tbody>
</table>

### 7.4.2 Calibration

Some parameters of a MATSim simulation can be calibrated to obtain more realistic results. In this work, there are two sets of parameters to calibrate. The first set is composed by the mode of transportation marginal utilities, and the second is composed of the frequencies of performing flexible activities.

#### 7.4.2.1 Mode share calibration

When agents schedule activities during the simulation they also decide which mode of transportation to use. The calibration of these parameters allow to match real mode shares with simulated mode shares. Figure 7.13 shows the histogram of departures in the relaxed simulation by all modes except car, compared with public transport departures recorded in CEPAS data. The number of departures in the simulation is multiplied by 10 to match with CEPAS transactions.

Although time distribution of departures are accurate on weekdays, the number of people traveling in the morning peak is half than the number of agents in the simulation. This is mainly due to the use of other modes in reality. In Singapore taxi is very popular, people use shuttle or company buses to go to work, students use private school buses, and some trips are done just by walking. Table 7.7 shows the description of the included modes and the calibrated parameters. In the section 7.5.3.2 results of this calibration are presented.
7.4.2.2 Flexible activity frequencies calibration

According to the function presented in Equation 4.3, the frequency of performing activities affect the decision of agents. If two activities of the same type are performed in a time shorter than a given period, the activity utility will be penalized. As no continuous multi-day data is available, observed frequency distributions of flexible activities can not be used in this model. Hence, as a first solution, one value of frequency was calibrated for each flexible activity type, to match observed activity performing shares by day of the week. Table 7.8 present the final values, and in Section 7.5.1, simulated activity shares are compared with observed shares in HITS.
Table 7.8: Calibrated flexible activity frequencies for the weekly simulation.

<table>
<thead>
<tr>
<th>Activity type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shopping</td>
<td>3 times per week</td>
</tr>
<tr>
<td>Eating</td>
<td>2.5 times per week</td>
</tr>
<tr>
<td>Running errands</td>
<td>0.4 times per week</td>
</tr>
<tr>
<td>Recreational activities</td>
<td>0.3 times per week</td>
</tr>
<tr>
<td>Medical activities</td>
<td>0.3 times per month</td>
</tr>
<tr>
<td>Religious activities</td>
<td>1.5 times per week</td>
</tr>
</tbody>
</table>

Figure 7.14: Computation time of the iterations of the weekly simulation

7.4.3 Simulation summary

The simulation of 10% of the population (371,996 agents) takes 1.5 hours for the first iteration and 30 minutes in average for the others as shown in Figure 7.14. Hence, to execute 100 iterations takes 2 days and 3 hours using a total of 60 GB of RAM. This statistics show that the process is computationally tractable.

Figure 7.15 shows that the average executed utility increased from 214 utils to 889 utils in only 20 iterations. This shows that routing on-the-fly using updated travel time is a very efficient strategy to find optimal activity-trip plans.

Figure 7.16 shows the car histograms of the weekly mobility simulation in iteration 0 and in iteration 20. The histogram of Cars en-route shows
that in just 20 iterations, agents found car routes and activity times which allow all agents to reach final destinations for every day in the week.
7.5  Results analysis and validation

7.5.1  Activity analysis

Fixed activities in this model are defined by the skeletons modeled in Chapter 5 and validated in Section 7.2. Therefore, this section focuses on the validation of flexible activity share. These results are based on synthetic Activity agendas, calibrated activity frequencies, and the way agents schedule activities in their free time. Figure 7 compares, for each day of the week, the reported flexible activity share in HITS with the share
of scheduled flexible activities in the simulation.

This model replicates that shopping is the most frequent activity in Singapore, and eating is the second. Activity shares on Monday are not as accurate as on Tuesday or Wednesday due to previous shopping activities on Sunday do not affect shopping performing on Monday. The model also replicates that shopping is not that popular on Fridays, and eating is performed more. Figure 9.4 also shows how agents increase the scheduling for dinner on that day. More religious activities on weekends are well modeled as well. However, the very low shopping rate on weekends has to be investigated. Next sections present how accurate flexible activities are scheduled in space and time.

### 7.5.2 Spatial analysis

In this section the spatial accuracy of simulated flexible activities is analyzed. Figure 7.18 compares the spatial distribution of shopping facilities (7.18(a)) with the distribution of reported shopping activities in HITS (7.18(b)), and with the distribution of simulated shopping activities (7.18(c)).

Results demonstrate that, although the distribution of the facilities is far from similar to the distribution of reported activities, the distribution of simulated activities is much closer. Hence, although agents can decide
Figure 7.18: Spatial comparison of shopping in Singapore.

(a) Shop facilities density (facilities/km²)
- 0.0 - 8.8
- 8.8 - 26.3
- 26.3 - 103.4
- 103.4 - 349.7
- 349.7 - 415.4

(b) HITS shop activity density (#activities/km²)
- 0.0 - 2.0
- 2.0 - 6.8
- 6.8 - 12.5
- 12.5 - 36.4
- 36.4 - 102.4

(c) SIM shop activity density (#activities/km²)
- 0.0 - 23.5
- 23.5 - 99.2
- 99.2 - 209.8
- 209.8 - 479.5
- 479.5 - 1084.5
Figure 7.19: Spatial comparison of medical activities in Singapore.
where to shop giving the supply of facilities (7.18(a)), they prefer to do it in facilities where real people shop. This is due to the distribution of home, work and study locations, to the utility maximization algorithm, and the selection of evoked places. For the case of shopping, the differentiation between SHOP HIGH and SHOP LOW place types was very successful. As some agents choose SHOP HIGH places, they have to travel longer from their home, work of study location, resulting in a more realistic spatial distribution.

In contrast, the simulated spatial distribution of medical activities (7.19(c)) is quite different to the reported distribution (7.19(b)). When analyzing the most popular areas in the simulation, it was found that the most popular HEALTH facilities in these areas are private hospitals and clinics. In reality, public hospitals are more popular, and they are located in the densest areas of Figure 7.19(b). If the evoked places model includes these two types of HEALTH locations the simulation error could be reduced.

Although results are not accurate, they show that the classification of locations is a key element in this model. Spatial distribution comparison of other activities can be found in section 9.3 of the Appendix.

### 7.5.3 Temporal analysis

As start time and duration of fixed activities never change during the simulation, the first part of the simulation’s temporal analysis focuses on flexible activities. The second part validates the simulation results with regards to agents using public transport against CEPAS smart card data.

#### 7.5.3.1 Flexible activities

Results show that flexible activities are scheduled according to start time distributions included in agents’ Activity agendas. Figure 9.4 compares the start time distributions of eating activities reported in HITS with the same distributions of simulated eating activities.

Breakfast, lunch and dinner peaks are well predicted everyday, including the popular Friday dinner. Saturday and Sunday lunches are slightly over-scheduled while Saturday and Sunday dinners are under-scheduled. This effect could be based on a different eating behaviour of people during weekends. People could sacrifice to travel for lunch to have a better dinner with family and friends. The model also generates a different pattern
Figure 7.20: Start time distributions of eating activities: HITS vs. Simulation

Compared with weekdays, but the change is not enough. A possible solution is to model different activity frequencies for different periods in the week. Other start time distribution comparisons can be found in Section 9.1 of the Appendix.

Figure 9.10 shows the comparison between the duration distribution of shopping activities reported in HITS and the corresponding simulation distributions.

As the reported duration of shopping activity is included in agents’ Activity agendas, duration of scheduled activities is predicted accurately. On weekends, shopping activities in the simulation are slightly longer. This might be caused by longer free time windows during these days. Comparisons of duration distribution of other flexible activities can be found in Section 9.2 of the Appendix.
Figure 7.21: Duration distributions of shopping activities: HITS vs. Simulation

7.5.3.2 Weekly public transport usage

In Figure 7.22, the weekly histogram of public transport departures in CEPAS data is compared with the augmented histogram of public transport departures in the simulation. The numbers of the simulation histogram are multiplied by ten to inflate the 10% scenario.

During the first days of the week the histograms are quite close to the CEPAS measurements. Morning and evening peaks happen at very similar times because they mostly depend on fixed activities, and times of fixed activities were synthesized from CEPAS data. In the simulation, the evening peak is less popular during the week, while it remains stable in reality. In between the peaks, the ascendant shape of the histogram is also replicated by the simulation during weekdays. Even more, on Friday, when less people go to work according to the clusters found in Chapter 5, the noon peak is reproduced in the simulation as well.

On weekends the simulation is not that accurate, many trips start earlier than in reality. Although Figure 7.4 shows that start time of simulated
home activities is well predicted on weekends, and Figure 7.5 shows that home activities are predicted to be longer on weekends, agents start flexible activities earlier than expected. This phenomenon shows up a flaw in this model, which is more evident on weekends. After assigning a skeleton to each weekday worker, the model assigns home activity times from time distributions of detected home activities. This is done for every day of the week. But there are home activities which take two or more days, and if they are detected in the public transport data, they are not included in time distributions of intermediate days. For example, when a person stays at home the whole Sunday, that activity is just included in the Saturday distribution, but not in the Sunday distribution of home activities. Then, random start time of home activities on Sunday are always going to be on Sunday evening and never on Saturday evening. In other words, this model
does not take into account that there are agents which stay the whole day at home during some days of the week.

7.6 Conclusions

This chapter summarizes how models developed in Chapter 5 and Chapter 6 can be used to set-up and execute the real large scale scenario of Singapore. New models to assign fixed activity skeletons to a synthetic population are introduced and validated. In the same manner, models to assign Activity agendas and Evoked places are explained and results are compared against the household travel survey of this city.

Using the synthetic population previously developed for daily simulations, skeletons, activity agendas and evoked places are assigned to a 10% sample (371,996 agents). Using the MATSim extension presented in Chapter 3 and the activity scheduling algorithm elaborated in Chapter 4, simulations with a week time horizon were executed.

Results show that start time and duration of flexible activities were efficiently replicated. For shopping, eating and religious activities, simulated spatial distributions were also accurate. Furthermore, scheduling activities on-the-fly with updated information produces a fast evolution of weekly agents’ utilities. In only 20 iterations, efficient car routes and activity times were calculated to allow all agents to reach their final destinations during the week with acceptable congestion measurements.

On the computation side, the iterative process needed a total 60GB. This information includes fixed activity skeletons, one weekly plan in memory per agent, flexible activity agendas, sets of evoked places and experienced travel times, among others. The simulation took 1.5 hours for the first iteration and 30 minutes in average for the others. In 2 days and 3 hours, 100 iterations were executed with agents closed to reach User Equilibrium.

Results also show that the model can be improved. Fixed activity skeletons were modeled with information of frequent public transport users, but fixed activity times were applied to all the agents. The assumption that the whole population behaves in the same manner as this sample is not validated. Therefore, the fixed activity times of non-public transport users may not be appropriate. Continuous multi-day information can be used to improve the model and to validate this assumption.
Agents who stay at home the whole day are not properly modeled because it is assumed that the home activity ends every day. The time when home activities end is modeled using joint distributions from frequent public transport users. People who stay at home are not present in this data, therefore, their time patterns are not included. As this is more common on weekends, fixed activities on these days show incorrect patterns.

The lack of continuous multi-day information affects the flexible activity models as well. Frequencies of performing flexible activities are unknown, and only one value per activity for the whole population was used. These values could only be calibrated with continuous multi-day data. Furthermore, the utility penalization for performing the same type of activity at higher frequencies could not be estimated with real data.

Finally, the spatial distribution of medical activities was far from close to the reported one. A classification of hospitals into public and private may help to solve this issue.
Chapter 8

Summary and Conclusions

8.1 Summary

This thesis proposes a model for large-scale multi-day mobility simulations based on activities. The previously developed multi-agent transport simulation platform MATSim was extended to execute and optimize the mobility of a population for horizons longer than one day. The concepts of fixed and flexible activities were introduced, allowing agents to start the simulation with incomplete plans of fixed activities and schedule flexible activities on-the-fly. This methodology allows to model activity type, time and destination choices according to current interactions with others, i.e. congestion, crowdedness, capacity of activity facilities, etc.

The design of this MATSim extension (Passive Planning) is based on enrichment of the population model, and the development of a new module. Two key elements in this new module are in charge of decoupling and synchronizing the mobility simulation and the planning processes: Planning engine and Planner manager. As planning processes are fully parallelizable, this approach allows to run in parallel any number of available threads for planning while the mobility simulation is executed. Two simple case studies were prepared to evaluate the efficiency of the new approach. The first successfully shows how agents can use updated information to improve their final utility. The second shows how agents react to an unexpected event in a multi-day scenario.

A new recursive multi-activity scheduling algorithm that doesn’t prioritize any scheduling dimension was designed and implemented. The utility-maximization algorithm returns the sequence of activities without imposing any size and the trips between them. For each activity within a time window, the method calculates start time, duration, location and type; and for each
trip the algorithm returns the travel time and the mode. Tests with controlled
inputs demonstrated the potential of the model. It is possible to calibrate the
number and duration of specific activities varying the corresponding utility
parameters. Spatial, temporal and tour restrictions can also be imposed.

This algorithm was also tested within the Passive Planning extension to
perform massive flexible activity scheduling tasks. Results show that the
process is computationally feasible, the first iteration (where all agents plan
flexible activities on-the-fly) took 3.5 minutes for a 1% sample (37,425
agents). This test used 20 threads for parallel flexible activities planning
and 30GB of RAM memory. Furthermore, when comparing a Passive
Planning process, i.e. planning flexible activities on-the-fly, with a standard
MATSim evolutionary process, the agents’ general utility increases more
than 50% more: Passive Planning improves the average utility from 100 to
140 utils. while the standard MATSim from 100 to 120 utils. This happens
because agents can schedule new flexible activities with new conditions of
travel times every iteration.

A weekly simulation of the previously developed Singapore scenario
was prepared and executed to evaluate this method with real data. For the
preparation, several models were developed to assign (i) a weekly fixed
activity skeleton, (ii) a flexible activity agenda and (iii) a set of evoked
places to each agent of the synthetic population. The main sources of
information to estimate these models were the household interview travel
survey (HITS), smart card transactional data (CEPAS), and a collected
database of activity facilities in Singapore. Thus, for each agent, the weekly
skeleton determines its initial incomplete plan, and the flexible activity
agenda together with the set of evoked places are used to schedule flexible
activities with the proposed recursive algorithm.

To assign weekly fixed activity skeletons, temporal weekly patterns of
primary activities performed by frequent public transport users in Singapore
were recognized from CEPAS transactions. The proposed method consists
of two steps. In the first, activities reached and left by public transport were
extracted from a 1% sample travel survey of Singapore. Then, to estimate
the likelihood of an activity to be HOME, WORK/STUDY or OTHER type
by its duration and start time, the extracted activity records were used to
develop two discrete choice models: one for workers and one for students.
In the second step, consistent activities were extracted from public transport
smart card transactions recorded during one week, and classified by type
applying the estimated discrete choice models. Then, to summarize weekly primary activity patterns, 14-dimension vectors composed by start time and duration of *WORK* or *STUDY* activities during continuous 7 days, were generated for each frequent public transport user. Finally, to recognize the most common travel behaviors, the *DBSCAN* clustering algorithm was applied to these vectors.

Results show that 5-weekday workers are the most representative group of public transport users. These people are represented by three large clusters with differences on the start time and duration of the recognized working activities. Users without recognized working activities during the week were represented by the 2nd largest cluster, and in contrast, users with recognized working activities everyday were represented by two smaller clusters. The third largest individual cluster represents frequent public transport users with recognized working activities every weekday except Friday. For students, some of the largest clusters represent the behavior of studying more than 8 hours during the 5 weekdays. These clusters differ due to the duration of the study activity on Friday. A cluster with 3% of the population represents students who study every weekday except on Friday. Half time students, and people studying 6 days were also represented by small clusters.

To assign *flexible activity agendas* and *sets of evoked places*, flexible activity patterns of this city where extracted, estimating 13 *Binary logistic regression* models from HITS. 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. Observed travel time distributions in HITS were compared with predicted travel time distributions for each flexible activity. It was found that, although travel time distributions generated using all evoked places contain longer travel times, travel time distributions generated using the scheduled evoked places are more accurate. This could model the idea that, although there are places in people’s choice set which are far from the current location, the selected places are closer.
When executing a 10% sample weekly simulation (371,996 agents), results show that start time and duration of flexible activities were efficiently replicated. For shopping, eating and religious activities, simulated spatial distributions were also accurate. Furthermore, scheduling activities on-the-fly with updated information produces a fast evolution of weekly agents’ utilities. In only 20 iterations, efficient car routes and activity times were calculated to allow all agents to reach their final destinations during the week with acceptable congestion measurements.

On the computation side, the iterative process needed a total 60GB. This information includes fixed activity skeletons, one weekly plan in memory per agent, flexible activity agendas, sets of evoked places and experienced travel times, among others. The simulation took 1.5 hours for the first iteration and 30 minutes in average for the others. In 2 days and 3 hours, 100 iterations were executed with agents closed to reach User Equilibrium.

This work contributes to the state of transport modeling, behavioral science and agent-based simulation in many ways. Activity scheduling decisions are modeled trying to emulate human decision processes. Although data-driven techniques could produce better prediction efficiency, their model structures are not designed for this purpose. Furthermore, models generated using data driven techniques are not easy to modify according to assumptions about the future for transport and land-use planning. Decisions modeled in this work take into account activity times, locations and even the way of reaching those locations with experienced travel times. Heterogeneity in people’s preferences is also included in the model by assigning preferred activities and locations according to reported preferences extracted from a travel survey. These preferences depend on socio-demographic and geographical characteristics. These types of models can be used to predict activity type, locations or times in other disciplines such as marketing, land use planning or sociology studies.

8.2 Discussion and outlook

8.2.1 Singapore data

The models for multi-day simulations presented in this thesis deal with the current situation of transportation data in Singapore. There is not a reliable survey reporting several days continuously, but based on public transport
data weekly patterns could be extracted. Although this situation seems very specific, many cities in the world have the same type of datasets.

Because of these limitations several aspects of the simulation could be improved. Fixed activity skeletons were modeled with information of frequent public transport users, but fixed activity times were applied to all the agents. The assumption that the whole population behaves in the same manner as this sample is not confirmed. Therefore, the fixed activity times of non-public transport users may not be realistic. The only way of validate this assumption is to obtain a sample of continuous multi-day information of the general population.

### 8.2.2 Fixed activities

The idea of differentiating fixed and flexible activities was introduced to reduce the complexity of multi-activity scheduling. However, in reality, this differentiation is blurred. There are activities which are scheduled previously but are flexible. There are also non-primary fixed activities which are more complex to schedule in the skeletons like medical appointments. With the proposed framework, it is possible to send agents to plan other fixed activities or re-plan activities in the skeleton before or during the simulation.

With the proposed approach, the time of fixed activities have to be estimated very accurately before the simulation starts. At that moment the duration of free time windows is defined, and only during these times the plans can be modified. An agent can improve its utility by extending the duration of fixed activities because the presented scheduler can allocate fixed activities in the beginning or in the end of the free time windows. However, these durations can not be reduced. Hence, fixed activities should be added to the skeletons with minimum durations, latest start times and/or earliest end times.

Furthermore, agents who stay at home the whole day are not properly modeled because it is assumed that the home activity ends every day. The time when home activities end is modeled using joint distributions from frequent public transport users. People who stay at home are not present in this data, therefore, their time patterns are not included. As this is more common on weekends, fixed activities on these days show incorrect patterns.
To solve this issue, a different approach could be developed for fixed activities. A model which predicts the existence of a main activity per day and estimates its start time and duration depending on sociodemographic and geographic variables can be developed. Then, this model can be applied before the simulation starts to every agent for each day of the week to create skeletons. It could also be applied during the simulation including updated information of previous experiences. However, new techniques should be developed to include the influence of main activities happening in past or future days. For instance, the duration of an educational activity on Friday can be affected if the student plans to go to school on Saturday.

8.2.3 Flexible activities

The lack of continuous multi-day information also affects the flexible activity models. Frequencies of performing flexible activities are unknown, and only one value per activity for the whole population was used. These values could only be calibrated with continuous multi-day data. Furthermore, the utility penalization for performing the same type of activity at higher frequencies could not be estimated with real data.

Spatial distributions of flexible activities are not accurate enough. The selection of evoked places is based on the type of the place, the number of opportunities, and the travel time from primary locations. For flexible activities with massive location opportunities such as shopping and eating, this method produce acceptable results. When a small number of locations are available for a certain activity, their popularity depends on other variables which were not included. Sometimes this popularity can be modeled as a type, and the proposed methods can be applied. For instance, a classification of hospitals into public and private may help to improve the spatial distribution of performing medical activities. But sometimes the popularity depends on continuous variables such as comfort, originality and/or level of innovation, and the type or the number of opportunities can not model that.
Bibliography


Stopher, P. R., K. Kockelman, S. Greaves and E. Clifford (2008) Reducing burden and sample sizes in multi-day household travel surveys, Transportation Research Record, 2064, 12–18.


Chapter 9
Appendix

9.1 Start time distributions comparison

Figure 9.1: Start time distributions of shopping: HITS vs. Simulation
Figure 9.2: Start time distributions of running errands: HITS vs. Simulation

Figure 9.3: Start time distributions of recreational activities: HITS vs. Simulation
9.1. Start time distributions comparison

Figure 9.4: Start time distributions of religious activities: HITS vs. Simulation

Figure 9.5: Start time distributions of medical activities: HITS vs. Simulation
9.2 Duration distributions comparison

Figure 9.6: Duration distributions of eating: HITS vs. Simulation
9.2. Duration distributions comparison

Figure 9.7: Duration distributions of running errands: HITS vs. Simulation

Figure 9.8: Duration distributions of recreational activities: HITS vs. Simulation
Figure 9.9: Duration distributions of religious activities: HITS vs. Simulation

Figure 9.10: Duration distributions of medical activities: HITS vs. Simulation
9.3 Spatial distribution

Figure 9.11: Spatial comparison of running errands in Singapore.
Figure 9.12: Spatial comparison of eating in Singapore.
Figure 9.13: Spatial comparison of recreation in Singapore.
Figure 9.14: Spatial comparison of religion in Singapore.

(a) Religion facilities density

(b) HITS religion activity density

(c) SIM religion activity density