DISS. ETH NO. 25885

SIMULATING ANNUAL LONG-DISTANCE TRAVEL DEMAND

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

MAXIM JANZEN

Diploma Mathematics and Diploma Computer Science, University of Bonn

born on 12.03.1984

citizen of
Germany

accepted on the recommendation of

Prof. Dr. Kay W. Axhausen, examiner
Prof. Dr. Theo A. Arentze, co-examiner

2019

Contents

Abstract **vii**

Zusammenfassung **ix**

1 Introduction **1**
1.1 Motivation 1
1.2 Research Gap 2
1.3 Research Objectives 3

2 Background on Long-Distance Travel Demand Models **5**
2.1 Capturing Long-Distance Travel 5
2.2 Activity-Based Models 6
2.3 Agent-based Models 7
2.4 Continuous Planning 9

3 Continuous Target-Based Activity Planning **11**
3.1 Behavioral Targets 12
3.2 Core Algorithm 13
3.3 Conclusion 15

4 Adapting C-TAP to Simulate Long-Distance Travel Demand **17**
4.1 Long-Distance Travel Demand Characteristics 17
4.2 Decision Making 20
4.3 Numerical Solver 24
4.4 Conclusion 27

5 Destination and Mode Choice **29**
5.1 Previous Work 29
5.2 Decision Variables 31
5.3 Modelling Destination and Mode Choice 36
5.4 Efficient Implementation 41

5.5 Discussion and Conclusion 45

6 Data Collection 47
6.1 Previous Work 48
6.2 Data Sources 53
6.3 Methodology 55
6.4 Analysis of the Data Sets 63
6.5 Limitations 71
6.6 Discussion 74
6.7 Conclusion 75

7 Population Synthesis 77
7.1 Synthetic Population Requirements 77
7.2 Limitations of known Approaches 78
7.3 Framework 79
7.4 Application 85
7.5 Discussion 89

8 Application 91
8.1 Scenario Configuration 91
8.2 Analysis 96
8.3 Conclusion 105

9 Implementation Details and Operational Challenges 107
9.1 Interfaces 107
9.2 Calibration and Validation 108
9.3 Scalability and Efficiency 109
9.4 Conclusion 114

10 Discussion and Outlook 115
10.1 Field of Application 115
10.2 Outlook 116
10.3 Contribution 117
10.4 Conclusions 118

References 121

List of Figures

4.1 Illustration of state values dynamics 21
4.2 Illustration of the Nelder-Mead Algorithm 25

5.1 Decision process of an agent 46

6.1 Mobile phone tower locations of the selected municipalities 57
6.2 Visualization of the tour reconstruction algorithm 60
6.3 Failed tour reconstructions 61
6.4 Cumulative frequency of tour distances 65
6.5 Destinations of Ile-de-France residents 66
6.6 Histogram of the number of LDF tours for mobile persons 69
6.7 Average LDF tour rates per mobile person per month . . . 69
6.8 Average LDF tour rates per capita by municipality size . . 71
6.9 Confidence Intervals of LDF tours per capita 72

7.1 Traditional population synthesis technique (IPF) 79
7.2 Illustration of Histogram Matching 81
7.3 Population Synthesis Framework 86
7.4 Duration density distribution for long-distance tours 89

8.1 Season function illustration 96
8.2 Analysis of duration distribution 98
8.3 Analysis of interval distribution 99
8.4 Analysis of number of tours by agent 99
8.5 Analysis of seasonal attractiveness 103

9.1 Runtimes of parallel computing approach 113

List of Tables

5.1 Impact of number of alternatives on optimality and run time 44

6.1 Annual long-distance tour frequencies: Other studies . . . 50
6.2 Number of tracked persons by size of municipality 59
6.3 Number of persons performing LDF tours 67
6.4 Average number of LDF tours per capita 70

7.1 Sample of a population synthesis 88

8.1 Destinations used in Ile de France scenario 93
8.2 Modes used in Ile de France scenario 94
8.3 Number of tours by purpose 97
8.4 National destinations accessed in Ile de France scenario . . 101
8.5 International destinations accessed in Ile de France scenario 102
8.6 Number of business activities by day-of-week 103
8.7 Modal Split in the scenario 105

9.1 Impact of population size 110
9.2 Impact of simulated time 110
9.3 Impact of Network size 111

Abstract

Research in the field of transport planning focuses today on urban mobility and, thus, on daily activities. Hence, most of the surveys, models and simulations aim to add valuable contributions to this topic. However, long-distance travel is growing, because travelling has become more accessible and cheaper recently. Nowadays, almost half of all vehicle miles travelled are generated by long-distance travel. Therefore, it is crucial to transport planners and policy makers. Nevertheless, so far only few surveys or models cover this field. One of the reasons is the challenging task to capture data describing long-distance travel. This thesis shows that big data can potentially help to overcome this burden using GSM data. Due to different challenges there has been also no effort to develop an agent-based simulation for long-distance travel. The major obstacle is the time horizon that such a simulation has to cover. Long-distance travel can not be analyzed focusing on a single day, because long-distance journeys usually consume more time and are also planned well in advance.

This thesis introduces an agent-based simulation covering long-distance and long-term travel demand and, thereby, closes the research gap described above. Furthermore, the thesis demonstrates that the simulation presented can adequately simulate large scale scenarios in reasonable time. The main value of such a simulation is support of policy makers for big infrastructural investments such as new bridges, tunnels or airports.

Zusammenfassung

Heutzutage konzentriert sich die Forschung in der Verkehrsplanung auf urbanen Verkehr und damit auf Aktivitäten des täglichen Lebens. Die meisten Umfragen, Modelle und Simulationen beschränken sich deswegen auf den Nahverkehr. Der Anteil des Fernverkehrs an der Gesamtmenge der Personenkilometer wächst jedoch stetig, da Reisen in den letzten Jahren immer günstiger und leichter zugänglich geworden ist. Verschiedene Studien belegen zudem, dass inzwischen fast die Hälfte aller Personenkilometer durch den Fernverkehr generiert wird. Aus diesem Grund ist der Fernverkehr insbesondere für Verkehrsplaner sowie politische Entscheidungsträger von eminent hoher Bedeutung. Dennoch decken bisher nur wenige Studien dieses Segment ab. Ursächlich dafür ist vor allem die anspruchsvolle Aufgabe die notwendigen Daten zu erfassen, die den Fernverkehr ausreichend beschreiben. In dieser Arbeit wird gezeigt, dass Big Data das Potenzial hat diese Aufgabe zu lösen. Aufgrund mehrerer Herausforderungen gibt es bis heute auch keine agentenbasierte Simulation für den Fernverkehr. Das Hauptproblem hierbei ist der zu simulierende Zeitraum, da Fernreisen nicht tagesbezogen analysiert werden können. In der Regel nehmen Fernreisen einen längeren Zeitraum in Anspruch und werden lange im Voraus geplant.

Diese Arbeit präsentiert eine agentenbasierte Simulation, die das Fernverkehrsverhalten über einen langen Zeitraum simuliert, und schliesst damit die oben genannte Forschungslücke. Die Arbeit zeigt weiterhin, dass die entwickelte Simulation grosse Szenarien adäquat und in angemessener Laufzeit simulieren kann. Der Wert einer solchen Simulation liegt insbesondere in der Unterstützung politischer Entscheidungsträger bei großen Infrastrukturinvestitionen wie beispielsweise neuen Brücken, Tunneln oder Flughäfen.

Acknowledgement

First and foremost, I want to thank my supervisor Prof. Dr. Kay W. Axhausen. He gave me the opportunity to work and study at the IVT. He guided me and pointed me in the right directions, but also gave me the freedom to develop my own ideas. I also want to thank my co-examiner Prof. Dr. Theo Arentze who agreed to supervise my thesis despite the fact that it is not part of his current research. I gratefully acknowledge the financial support of the Swiss National Science Foundation. Furthermore, the collaboration with Orange Labs was an important part of my studies. Especially, Maarten and Aare ensured that the collaboration was very fruitful and that I could maximize the utility of my stay at the labs.

No acknowledgement would be complete without mentioning my colleagues at IVT. All of them made IVT a productive workplace that I will miss. Special thanks go to my office mates who shared the office with me from the very first to the very last day. We had many gainful discussions in the office and at various other occasions.

My parents and grandparents receive my deepest gratitude, without whom I would never have enjoyed so many opportunities. Last but definitely not least, I have to thank Elena who had to suffer from my idea to go to Zurich for a PhD. Your patience, encouragement and belief carried me through the last five years. You supported me all the time. Thank you for your love!

Chapter 1

Introduction

1.1 Motivation

To date, travel demand models focus on reproducing and predicting daily behavior. This stands in contrast to the significant part of traffic volume caused by long-distance journeys related to activities not usually undertaken during daily life. Various studies have reported that long-distance journeys account for almost half of all vehicle miles travelled. For example, Grimal (2010) reported a share of 40% for 2008 in France. Frick *et al.* (2014) showed that 45% of vehicle miles travelled in Germany 2014 were part of a long-distance journey. This share will potentially grow in the future since long-distance travel is getting cheaper and more people get access to the relevant travel modes. This is supported by the fact that there is still unmet need for long-distance travel (Ullman and Aultman-Hall, 2018). Therefore, analysis and forecast of long-distance travel behavior has become more important recently.

However, research on this topic has been limited so far as it will be shown in Chapter 2. Yet, a model for long-distance travel demand is valuable, because it introduces a new possibility to evaluate political decisions in this policy domain. An application might be the evaluation of big infrastructural investments, like new bridges or tunnels, which is beneficial for the cost-benefit analysis of these investments. Additionally, results for long-distance travel demand can be combined with short-term traffic simulations to get a complete image of total demand for travel.

1.2 Research Gap

Agent-based simulations have a long tradition in analysis and explanation of social behavior and were also used to estimate travel demand (Axhausen and Herz, 1989; Pendyala *et al.*, 1997) or to generate an activity-based travel forecast (Bhat *et al.*, 2003; Miller, 1996). Nowadays, agent-based simulations make a notable contribution to the field of transportation research (Balmer, 2007; Arentze *et al.*, 2010; Kuhnimhof and Gringmuth, 2009; Erath *et al.*, 2012). Microscopic travel demand simulations simulate the (travel) behavior of virtual agents individually. One of the well known approaches is the one proposed by Balmer (2007): agents choose a daily schedule for their behavior and execute it. The execution results are reported and the agents can re-plan their schedule based on the results of all agents. This procedure is iterated until a stochastic user equilibrium with consistent travel demand is reached (Nagel and Flötteröd, 2009). Due to high computational complexity and memory issues (all current schedules have to be maintained) a reasonable simulated period is a single day.

Long-distance trips have been the focus of recent literature. Long-distance travel behavior has been analyzed several times, e.g. for the UK and the Netherlands (Limtanakool *et al.*, 2006). Statistical long-distance travel demand models were developed in Erhardt *et al.* (2007) and used for traffic forecast (Beser and Algers, 2001). Recently, different surveys were analyzed to derive an outlook on the future of long-distance travel demand (Frick and Grimm, 2014; Outwater *et al.*, 2015a).

However, as of yet agent-based simulations have not been utilized to predict long-distance travel demand. The main reason is the iterative equilibrium approach that limits the time horizon of most frameworks. Arentze and Timmermans (2006) introduced an alternative approach, the need-based theory, which allows continuous activity planning. It was further modified by using behavioral targets leading to a continuous target-based approach by Märki (2014). Agent-based simulations are beneficial, because statistical models of long-distance travel behavior focus on the current state of the world, which is not always sufficient. Thus, there is a need for a tool to predict travel demand after major infrastructural or cost changes which introduces a new possibility to justify political decisions in this policy domain. Further advantages of agent-based modelling have been summarized by Helbing and Balietti (2012). It provides flexibility

(definition of behavior is flexible), heterogeneity (individual variations are possible) and stochasticity (random variations can be implemented). The target-based simulation presented in this dissertation is the first agent-based simulation which provides an estimate of long-distance travel demand.

1.3 Research Objectives

The main goal of this dissertation is the development and implementation of an agent-based simulation that generates long-distance travel demand for at least one year. The simulation has to accurately model travel behavior of a population as well as predict changes in travel behavior. Secondary goals are a simple interface and a reasonable simulation runtime. The core ideas ensuring these goals are pointed out in the following.

First, the microsimulation implemented has an open time horizon, i.e. it generates arbitrarily long activity plans. The simulation run time has to scale linearly with the time horizon. The main idea to ensure linear scaling is the implementation of continuous planning. Continuous planning carries out a concept of decision making without replanning steps. In other words, a decision made by an agent is irreversible. Continuous planning also does not involve any learning process. One of the advantages is the simple agent structure that is needed. There is no need of maintaining a set of plans, or all activity history. An agent is rather defined by small set of attributes, so called *targets*. This approach saves memory and scales the memory needed linearly with population size. As a consequence, an increase of the time horizon does not require additional memory (RAM). Therefore, the length of the generated activity plans can be of arbitrary length.

A further important requirement of the agent-based simulation is reasonable runtime paired with scalability. In other words, the runtime should not grow exponentially with growing population size or an increase of the time horizon. Scalability is achieved by the activity planning procedure which is continuous and implies simple agent structure. A further step to establish a micro-simulation with adequate runtime is parallelization of the software. As it will be shown in Chapter 3 and Chapter 4 the decision process of each agent can be parallelized easily. Chapter 4 also describes the decision making process and the heuristic used to ensure acceptable simulation runtimes.

As mentioned above, a realistic simulation output is crucial for the value of a travel demand simulation. More precise, the simulation has to be able to reflect observed behavior. In a second step, it has to predict behavioral changes that arise from infrastructural changes. This task can be split in two sub-tasks. First, the simulation itself has to turn input data that describes the real world into reasonable output data. In our case, input data is mainly given by a virtual population and a network and simulation output is an activity plan for every agent modelled. The development of the simulation is described in Chapter 4 and Chapter 5. Second, the configuration of the simulation is essential. In our case, the main configuration part is a population of virtual agents reflecting the real world. A valuable virtual population demands reliable data collection as presented in Chapter 6 as well as a population synthesis that scales the collected data to a full population as described in Chapter 7. Finally, Chapter 8 shows an application of the simulation using the virtual population as described in Chapter 7.

Chapter 2

Background on Long-Distance Travel Demand Models

The aim of this work is the implementation of a continuous agent-based simulation of long-distance travel demand. Previously, agent-based models have not been implemented for this purpose. This dissertation is a first step to bridge this research gap. Different streams of research had an impact on this work. The relevant literature is summarized in the remainder of this chapter. First, an overview of the problems that arise when collecting long-distance travel data is given. Second, the first modelling approaches, namely activity-based econometric models, are recapped. Then, agent-based modelling and its traditional iterative implementations are presented. Finally, need-based theory which allows continuous planning is introduced.

2.1 Capturing Long-Distance Travel

Models of travel demand require data sources that capture actual travel or at least a sample of it. Long-distance travel is no exception to this rule. Nevertheless, collecting long-distance travel data has been proven to be challenging.

The first challenge is the definition of long-distance travel which is inconsistent in literature. A trip distance threshold is widely used to separate long-distance travel from the remainder. However, there are variations in two different dimensions. First, the threshold distance differs in the studies. Thresholds used include, but likely are not limited to, 50 km (e.g. Limtanakool *et al.*, 2006), 80 km (e.g. Armoogum *et al.*, 2008), 100 km (e.g. Chlond *et al.*, 2006), 50 miles (e.g. Bierce and Kurth, 2014) and

100 miles (e.g. LaMondia *et al.*, 2008). Second, there are two different ways to implement the distance threshold, either as crow-fly distance or as actual distance travelled. The latter option has been used more frequently since recalling this value requires less effort for survey respondents. Furthermore, some studies consider overnight stays as long-distance travel regardless of the distance travelled, e.g. Armoogum *et al.* (2008).

Due to the amount of variations in the definition of long-distance travel a consistent comparison of relevant measures (like the share of miles travelled) is challenging. Moreover, the response burden of long-distance surveys yields inaccurate data. Various studies and issues of capturing long-distance travel were summarized by Axhausen *et al.* (2002a).

Single surveys on long-distance travel demand have limitations. Due to a high response burden the number of long-distance tours reported is usually far below the actual long-distance travel. Chapter 6 provides a deeper insight into this issue. The most prominent approach to overcome the problems of a high response burden has been a combination of several surveys which focus on separate types of long-distance travel. A combination of surveys has been done successfully in Germany (Kuhnimhof and Last, 2009; Frick and Grimm, 2014) as well as in California (Bierce and Kurth, 2014; Cambridge Systematics Inc., 2013).

2.2 Activity-Based Models

Collecting data has been the first obstacle for predictions of long-distance travel. In a second step, reliable models are needed. Trip-based models have been the simplest version of a statistical model for travel demand. This has been also the case for models of long-distance travel. Trip-based models mostly ignore the cause of travel, namely activities. One of the reasons for this approach is the lack of information about travel purposes. This is the case if the source of the underlying data is passive data collection like GPS or GSM data. For example, Bekhor *et al.* (2013) took mobile phone data plus a national travel survey to reveal long-distance travel patterns on a trip basis.

However, a crucial assumption in travel demand modelling is that all travel is induced by a desire to carry out certain activities at certain locations (e.g. Chapin, 1974; Axhausen and Gärling, 1992). This assumption applies

also to long-distance travel modelling and have resulted in the state of the art: activity-based models. Activity-based models explain all travel as a result of activities at different locations.

Activity-based models have been used to explain all pieces of long-distance travel demand. Activities in these models usually induce home-based tours. For example, binary choice models for long-distance tours were used in Ohio (Erhardt *et al.*, 2007) to explain the motivations for long-distance travel. In a next step, mode choice models instantiating nested-logit models have been implemented on several occasions in Europe (e.g. Rich and Mabit, 2012) or USA (e.g. Koppelman and Sethi, 2005; Moeckel *et al.*, 2015). In California, Structural Equations Models (SEM) as well as Latent Class Analysis was performed to estimate miles travelled by mode and the length of long-distance tours (Goulias *et al.*, 2018).

Adding all activity-based models together, one can generate a full picture of long-distance travel demand. For instance, a sequence of interdependent sub-models representing choice of tour frequency, tour destination, travel mode and other related choices (Koppelman, 1989). This approach has been further enhanced to develop a tour-based national model in the USA (Outwater *et al.*, 2015a; Bradley *et al.*, 2015).

Nevertheless, econometric models as those presented so far have limitations. One of them is the lack of model heterogeneity, i.e. large variations in behavior. Agent-based models have been always a proven method to overcome this drawback and add more flexibility to the modelling.

2.3 Agent-based Models

A special case of activity-based model are agent-based models. The core idea of an agent-based model is disaggregation of travel demand modelling. The population of the real world is modelled by a virtual population consisting of agents. Each virtual agent represents a real person and reproduces his or her travel behavior. Agent-based travel demand simulations simulate the (travelling) behavior of virtual agents individually.

Advantages of agent-based simulations for simulating human behavior were summarized by Bonabeau (2002). First, agent-based models are able to capture emergent phenomena, i.e. simulation of the behavior of the system's constituent units (the agents) and their interactions captures

emergence from the bottom up when the simulation is run. Second, agent-based modelling provides a natural description of a system, because each person and its behavior is modelled individually by a single agent. Finally, agent-based modelling is flexible since the modeller can easily add or change agents in the system without manipulating the core of the simulation. These characteristics allow to simulate changes in the system and effects triggered by these changes. This is beneficial, in particular for analysis of future infrastructural changes.

One of the well known approaches for agent-based simulations of travel behavior is MATSim (Balmer, 2007; Balmer *et al.*, 2006; Horni *et al.*, 2016). In MATSim agents choose a daily schedule for their behavior and execute it. The execution results are reported and the agents can re-plan their schedule based on the results of all agents. This procedure is iterated until a stochastic user equilibrium with consistent travel demand is reached (Nagel and Flötteröd, 2009). Further equilibrium based simulations were implemented earlier, e.g. METROPOLIS (de Palma *et al.*, 1997) or TRANSIMS (Smith *et al.*, 1995). Agent-based simulations without equilibrium seeking were developed with an emphasis on specific research tasks. For instance, AgentPolis (Jakob and Moler, 2013) focuses on journey planning in a multi-modal network which is beneficial for analysis of the impact of arising travel modes in an urban environment. Axhausen and Herz (1989) were focusing on the simulation of pure activity chains.

An alternative to traditional agent-based equilibrium has been developed by calculating a dynamic network equilibrium on a vehicle basis (Halat *et al.*, 2016). However, due to high computational complexity and memory issues (e.g. all current schedules have to be maintained) a reasonable simulated period is a single day. The same holds for all agent-based simulations introduced above. Simulating a single day is not sufficient to describe long-distance travel demand since long-distance travel is not planned and executed on a daily basis.

No alternative agent-based simulations have been proposed to simulate long-distance travel demand. The main limitation of traditional agent-based models has been the time horizon of a single day which is not sufficient for modelling long-distance travel demand. Long-distance travel demand needs an activity planning horizon that goes beyond a single day. An idea of longer simulation periods have been pushed recently. For example, Ordóñez Medina (2017) modified the MATSim model in order to simulate

travel behavior of a week. Despite that, due to a lack of an appropriate tool long-distance travel demand has not been subject of agent-based simulations. The tool used in this work in order to develop an agent-based simulation of long-distance travel demand is the need-based theory which is introduced in the following section.

2.4 Continuous Planning

The target-based approach presented here is related to the need-based theory which was introduced by Arentze and Timmermans (2006, 2009). They developed a model for activity generation with the assumption of utilities described as dynamic function of needs. This was utilized in a rule-based approach to dynamically generate activity agendas on a household level. The agendas were generated with respect to time-budgets, perception, altruism and joint activity participation. Nijland *et al.* (2014) extended the need-based model employing a decision heuristic to account for future events and conditions.

The idea of needs suits the idea of continuous planning as shown by Märki *et al.* (2011) which is crucial for a simulation with an open time horizon. However, it is unclear which needs trigger which activities. Additionally, a need-based model requires estimation of many parameters. Therefore, targets instead of needs were used as an explanation of human behavior (Märki *et al.*, 2012a). Targets have the advantage of a direct association between model parameters and their induced behavior.

Behavioral targets, follow a similar idea as needs. Agents are defined by their behavioral targets. They satisfy these targets by executing corresponding activities. Therefore, one has a direct link between activity motivation and activity execution which makes targets more intuitive than needs. In contrast to the need-based approach, a user can provide a direct definition of the targeted performance. For instance, a duration target for certain activity defines the duration of an activity that is desired by the agent. Various types of targets can be defined to induce corresponding behavior, e.g. percentage-of-time targets or frequency targets. The target-based approach has been validated for daily life travel demand generation. Märki *et al.* (2012b) were able to reproduce travel behavior of a six week household travel diary from Switzerland. Finally, a travel demand generation framework that allows for

a continuous simulation of large scale multi-week scenarios was presented and validated (Märki, 2014; Märki *et al.*, 2014a). The main ideas and the core algorithm of the continuous target-based activity planning are introduced in the following chapter.

Chapter 3

Continuous Target-Based Activity Planning

The main idea of an agent-based simulation is that a virtual population makes decisions on its behavior and generates individual activity plans. Each of these plans includes activities, locations, travel modes, departure times, routes used etc. Execution of each of these plans yields effects on the environment such as congestion or emissions. These effects can be taken for analysis of infrastructural changes.

In contrast to iteration-based models (like the one used by Balmer (2007)) a continuous planning model does not iterate to a steady state, but generates continuously an activity schedule without any replaning. Iterative approaches have a fixed activity sequence and optimize the whole activity plan at once. In case of continuous target-based planning, the planning is dynamic, i.e. no complete schedules are created. The sequence and scheduling of activities is not defined a priori. For each agent a decision about his next activity has to be made each time he finishes his current activity. One of the main advantages is the ability of the simulation to generate arbitrary long activity plans in linear runtime (linear in simulated time). Thus, it is more useful to generate long-term travel demands.

The **C**ontinuous **T**arget-based **A**ctivity **P**lanning (C-TAP) simulation model presented in this chapter was introduced by Märki *et al.* (2012a), further developed in Märki *et al.* (2014b, 2013) and summarized in Märki (2014). The main ideas of C-TAP, namely the targets, activities and their interaction in the algorithm are recapped in the remainder of this chapter.

3.1 Behavioral Targets

Agents in C-TAP are built of behavioral targets. Targets can be seen as motivations of an agent to execute certain activities. There are several options to define targets. In the following the three types of targets proposed by Märki are presented:

- percentage-of-time target: indicates how much relative time within an observation window an agent would like to spend on a specific activity.
- frequency target: indicates how often an agent would like to execute a specific activity within an observation window.
- duration target: indicates how much time an agent would like to spend for a single execution of a specific activity.

Note that the first two target types include a definition of an observation window. It is worth mentioning that any two of the three target types are sufficient to describe the motivation of the agents. This redundancy is utilized in a later chapter.

3.1.1 Activities and State Values

Activities are necessary to complete the concept of a target-based simulation, because the targets (or motivations) described above are satisfied by the execution of a corresponding activity. Accordingly, activities also mark travel purposes. The decision on the execution of activities is based on state values. A state value is defined for each target which is necessary to measure the satisfaction of targets. Two types of state values are needed:

- for targets with observation windows (percentage-of-time and frequency targets): the state value is an exponentially weighted moving average which is restricted to the length of the observation window. It increases during the execution of the relevant activity, respectively decreases during non-execution.
- for duration targets: the state value is defined as the activity duration.

The state values for targets with an observation window is calculated as a convolution of the activity execution pattern with an exponential kernel which is restricted to the length of the observation window. The level of satisfaction is measured by the quadratic difference of state value and target value. This measurement is called *discomfort* and its influence within the

model is described in detail in the following section.

3.2 Core Algorithm

The core algorithm of the simulation can be summarized in an event-based loop as it is shown in Algorithm 3.1 in pseudo code.

Algorithm 3.1: Core C-TAP Algorithm (Pseudo Code)

```
while simulation end not reached do
    for all agent with no activity do
        state ← UpdateAgentState(agent)
        nextActivity ← MakeDecision(agent, state)
        agent.execute(nextActivity)
    end for
    nextTimeStep = minimum( all execution endpoints)
    proceed to nextTimeStep
end while
```

The main procedure is a continuous iteration over points of time. Every time an agent finishes the execution of an activity the function $MakeDecision$ computes the next activity based on its current state. After that, the activity is executed until the computed execution ends. Activity execution also includes traveling to the location of the activity. Recording this trips generates the travel demand. The simulation stops after a predefined stopping condition is reached. This condition is usually a time period, which has to be simulated. The traditional C-TAP implementation as presented in Märki (2014) simulated several weeks of travel behavior. In case of long-term simulations, which is the goal of this work, a time period of one year is reasonable. The implementation of the $MakeDecision$ function is the crucial part of the simulation since it specifies the activity planning. The activity planning function utilizes the target-based theory and is the main topic of the following section.

3.2.1 Decision Procedure

The decision process of the agents in C-TAP is driven by a measure of (target) satisfaction, namely the discomfort value. The discomfort D_α of an execution of activity α is the sum of quadratic differences between targets and their corresponding state values:

$$D_\alpha(t) = \sum_{k=1}^{n_\alpha} (f^k_{target}(t) - f^k_{state}(t))^2 * w, \tag{3.1}$$

where n_α is the number of targets connected to the activity considered and w a bandwidth normalization factor. The function $f^k_{target}(t)$ describes the target value of a given point of time t, while $f^k_{state}(t)$ describes the state value at t. The discomfort function is used to define the discomfort reduction DR_α of an activity execution:

$$DR_\alpha(t^s, t^e) = D_\alpha(t^s) - D_(t^e), \tag{3.2}$$

where t^s is the start time of the activity execution and t^e is the end time. The discomfort reduction function gives an idea how an activity potentially improves the satisfaction of an agent.

The decision procedure is a sequential process. The first step is to maximize the product of $DR_\alpha(t^s, t^e) \cdot LA_\alpha(t^e)$ for every available activity. More precisely, the t^e which maximizes the product has to be found. LA_α is a look-ahead function which is used to measure future execution options. Märki proposes the Brent optimization algorithm (Press *et al.*, 2007) to solve this maximization problem. The next step consists of multiplying the maximized product with an execution time ratio $\frac{t^e - t^s}{t^e - t^{ts}}$ and an optional random term $(1 + \epsilon)$, where t^{ts} is the travel starting time. This results in a heuristic function

$$HF_\alpha(t^{ts}, t^s, t^e) = DR_\alpha(t^s, t^e) \cdot LA_\alpha(t^e) \cdot \frac{t^e - t^s}{t^e - t^{ts}} \cdot (1 + \epsilon). \tag{3.3}$$

After computing all HF_α-values for all considered activities the agent executes the activity, which yields the highest HF_α-value.

3.3 Conclusion

Activity planning generated by the target-based approach yields activity schedules which do not constitute a global optimum or any equilibrium. Each agent chooses a plan that appears to be the best considering the information available. However, this approach is valid for a simulation of long-distance travel demand. A long-distance journey is usually a rare event. Therefore, the travellers can not build on their daily experience, but decide based on their limited information.

Chapter 4

Adapting C-TAP to Simulate Long-Distance Travel Demand

C-TAP as it was initially introduced by Märki (2014) was implemented to simulate daily life travel behavior for multiple weeks. This initial implementation is referred to as traditional C-TAP in the following. Traditional C-TAP successfully reproduced six weeks of travel behavior from a household travel diary. The scope of this work is a larger scale, both in time simulated and trip distance travelled. The increase in scale necessitate less granularity for the simulation as it is shown in the remainder of this chapter.

This chapter is structured as follows: First, the modifications needed in order to simulate long-distance travel instead of daily travel are described. This is followed by an introduction of a modified decision making approach. Finally, a numerical solver for the new activity planning problem is presented. The findings presented in this chapter build on work previously presented (Janzen *et al.*, 2014; Janzen and Axhausen, 2015a,b).

4.1 Long-Distance Travel Demand Characteristics

The first implementation of C-TAP simulates daily-life activity planning over several weeks. The task of the implementation presented in this work, namely simulation of long-distance travel, differs substantially. The time horizon considered exceeds a few weeks since long-distance travel is relatively rare and usually planned in advance. Therefore, a reasonable time period simulated is a year. All adaptions that allow C-TAP to simulate

long-distance travel demand for a year are the topic in the remainder of this section.

4.1.1 Activities

The main objective of this work is the generation of long distance travel demand. Thus, the interest is not in every short trip, but just in those trips with long distances. C-TAP is an activity based model, i.e. the activities performed during a C-TAP simulation serve as travel purposes. The modification of traditional C-TAP to a long-distance travel demand simulator demands a transition to aggregated activities. For example, a single activity representing daily life which includes all short daily journeys like traveling to work, shopping, etc. is introduced. In comparison to Märki (2014) this is a higher abstraction level of activities.

A traditional segregation of long-distance travel purposes is business travel, vacation, private travel and commuting. For example, these purposes are implemented in an European travel demand model (Rich and Mabit, 2012) where the authors utilized the DATELINE survey (Brög *et al.*, 2003). The long-distance C-TAP simulation uses similar travel purposes with marginal modifications. Business travel and vacations are regular activities in C-TAP. Private travel usually covers the largest share of long-distance travel purposes. However, visits of family and friends takes a substantial part of this share, e.g. about half of all private long-distance tours were visits in an analysis of long-distance travel presented by Bricka (1999). Therefore, the private travel class is split into two separate travel purpose classes, visits and other private travels. The latter includes mainly recreation or shopping activities.

On the other hand, commuting is identified as a major long-distance travel purpose, but is left out of C-TAP. The reason is that C-TAP is a long-term activity planning tool simulating travel behavior outside of daily life. Commuting is part of daily life with a specific fixed schedule and does not need to be considered within long-term activity planning. Thus, long-distance commuting is not part of the C-TAP simulation modified for long-distance travel demand simulation.

4.1.2 Destinations and Modes

Travel modes considered for long-distance travel are different to the ones used in daily-life travels. Firstly, biking and walking is not a valid option for long distances. Secondly, public transport is available, but busses and LRT within city boundaries are replaced by long-distance heavy rail. Rail has a different price structure and travel speed than urban public transport. In addition, air travel is relevant for long-distance travel. Planes have characteristics of traditional public transport, e.g. a person does not need to own a vehicle and it is accessible to everybody. However, air travel also differs to public transport. The headway is higher and the price structure differs as it is dependent on variables such as capacities. Furthermore, access and egress time to airports is usually relatively high. The last mode considered, the car, has comparable characteristics to the corresponding daily used mode. The price (per km) is similar to daily life usage. Travel speed is usually higher, but is still linked to speed limits. Therefore, the three modes considered in the long-distance version C-TAP are rail, air and car.

Travel destinations in daily life are not fixed for secondary activities. Usually only locations for work and education are fixed. Destinations for secondary activities can vary, but are limited to specific facilities. The choice set for these facilities is limited to all facilities within a relatively small area. The area considered is usually the surroundings of the previous and the following activity and the area between these two activities (e.g. Horni *et al.*, 2012). This approach is not applicable to long-distance travel destination choice. Any activity beside the home activity can take place anywhere around the world. The level of destinations of interest in case of long-distance travel are municipalities instead of specific facilities. However, the number of potential destinations is too large to simulate a detailed destination choice. Therefore, a hierarchical approach is needed. The activity planning module in C-TAP can estimate the destination regions or countries. A detailed choice can be specified subsequently.

Implementation details of destination choice as well as mode choice within C-TAP are given in Chapter 5. An application of C-TAP with mode choice and destination choice on the top level of the destination hierarchy is shown in Chapter 8.

4.1.3 Extending the Planning Horizon

Traditional C-TAP is an activity planning tool that continuously makes decisions on the next activity performed. Heuristic approaches have been utilized to foresee certain conflicts beyond the single activity planned (e.g. due to shop closing times). However, one of the main modifications towards long-distance C-TAP is the extension of the planning horizon. In other words agents in the simulation plan more than one activity in advance. This is reasonable, because long-distance journeys are usually planned in advance. The advantage of an extension of the planning horizon is the expendability of a look-ahead heuristic since agents consider the outcome of a sequence of activities.

In contrast to the traditional C-TAP model, the future state values f_{state} are not dependent on a single activity, but on a sequence of activities. Thus, discomfort reduction is not a reasonable driver of the activity planning since it is limited to the impact of a single activity. The decision making within the activity planning has to be adapted to account for this change.

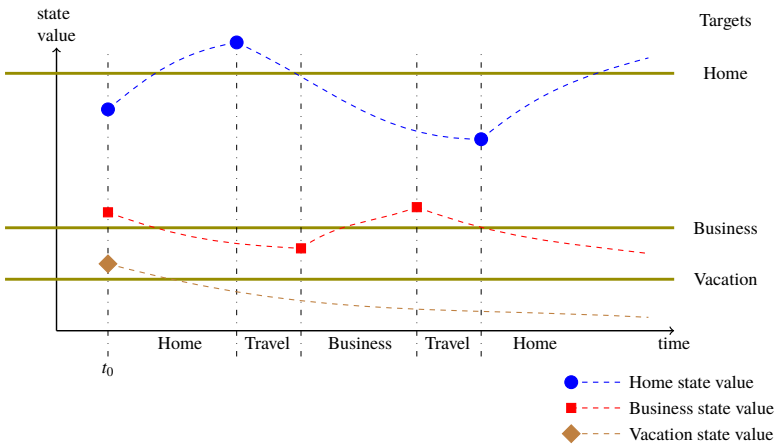
4.2 Decision Making

The definition of all three targets mentioned in Section 3.1 for a single activity has redundancy as any two of the three target types are sufficient to fully describe the motivation of the agents for a single activity. In order to reduce runtime and complexity of the model only two of the three proposed targets in Märki (2014) are implemented, namely the duration and percentage-of-time targets. These two target types are sufficient to describe long-distance travel demand. The frequency of an activity is modelled implicitly with these two targets. Furthermore, frequency of an activity is not intuitive for long-distance travel purposes, because certain frequencies are rarely targeted in this scope.

As described in the previous section, the activity planning module considers several activities in advance. This approach is not compatible with the idea of discomfort reduction maximization. Discomfort reduction takes into account only a single activity including its targets and state values. This is a limited view on the state of an agent. In order to get a full picture the full discomfort of an agent is considered in the decision planning. The full discomfort is the sum of all discomfort values of an agent. Thus, all

targets and all state values of an agent are considered at once. In other words, the discomfort reduction induced by an activity execution is taken into account as well as the discomfort increase induced by non-execution of other activities. The discomfort increases since percentage-of-time state values decrease any time the corresponding activity is not executed.

Figure 4.1: Illustration of state values dynamics



An illustration of the percentage-of-time state values is shown in Figure 4.1. Three percentage-of-time state values of an agent are sketched. The thick lines represent the percentage-of time targets of the three activities home, business and vacation. The x-axis shows the activity plan of the agent who performs home-business-home including travelling between the activities. One can see that the vacation state value is constantly decreasing due to the fact that vacation is never executed. The home state value increases twice during the execution of the home activity and decreases otherwise during travel and during business execution. Similar holds for the business state value. The full discomfort describes the sum of all discomfort values at the end of the planning horizon and is the function that is optimized in the activity planning.

4.2.1 Full Discomfort

The decision making in the long-distance version of C-TAP differs from the traditional C-TAP approach. Instead of computing a heuristic value which is based on the discomfort reduction of a single activity the full discomfort value is used as driver for the activity planning. The main difference is that all activities and their targets are considered at the same time. Furthermore, a sequence of activities is considered when a decision needs to be made. The length of the sequence is called planning horizon.

Following the description above, it is assumed that for every activity α there exist two targets, namely a duration target $T^{\alpha}_{dur} \in \mathbb{R}^{+}$ and a percentage-of-time target $T^{\alpha}_{pot} \in [0, 1]$. The discomfort resulting from the percentage-of-time targets depends on the definition of the corresponding state values. As discussed above these values are dynamic in the simulation. Hence, they are defined as a function of time t, denoted as $v^{\alpha}(t)$. The definition of v^{α}-functions explains the evolution of state values during the execution of activities. The state values are described by two exponential functions. First, there is a state value increasing function:

$$\hat{v}_{\alpha}(t_1, t_2, v) = 1 + (v - 1)e^{-\tau_{\alpha} \cdot (t_2 - t_1)}. \tag{4.1}$$

Second, a state value decreasing function is defined as well:

$$\check{v}_{\alpha}(t_1, t_2, v) = v \cdot e^{-\sigma_{\alpha} \cdot (t_2 - t_1)}. \tag{4.2}$$

In both cases v is the state value at the beginning, i.e. at point of time t_1. of the state value applies. τ_{α} and σ_{α} are constants that are computed for every activity individually and are subject to calibration. Note that valid state values v are between 0 and 1. Whenever an activity α is executed from t_1 till t_2, the corresponding state value increases to $\hat{v}_{\alpha}(t_1, t_2, v)$. Whenever an activity is not performed from t_1 till t_2 (the agent either travels or performs another activity), its state value decreases to $\check{v}_{\alpha}(t_1, t_2, v)$. Note that $\hat{v}_{\alpha}$ is a convolution of the state value function with an exponential kernel. Thus, it is additive in duration, i.e. $\hat{v}_{\alpha}(t_2, t_3, \hat{v}_{\alpha}(t_1, t_2, v)) = \hat{v}_{\alpha}(t_1, t_3, v))$. The same applies to the $\check{v}_{\alpha}$-function. This property simplifies computation the discomfort.

The percentage-of-time discomfort function D_{pot} after an execution of a sequence of n activities can be phrased as follows:

$$D_{pot}(\alpha, t^s, t^e) = \sum_{\omega \in \mathcal{A}} (T^{\omega}_{pot} - v_{\omega}(t^e_n))^2 \tag{4.3}$$

The function parameters are vectors of n dimensions describing n activities α_i and their corresponding vectors of start times t^s and end times t^e. Here, the end time of the last activity t^e_n marks the end of the planning horizon. Thus, it is the relevant point of time to measure the percentage-of-time discomfort. The set of all activities is denoted as $\mathcal{A}$.

As mentioned above, the state values v_α are dynamic and need to be calculated recursively following the definitions of $\check{v}_\alpha$ and $\hat{v}_\alpha$. Assuming a planning horizon of two activities α, t^s and t^e are two dimensional vectors. The evolution of state values for the two dimensional case ($n = 2$) is expressed as follows:

$$v_\omega(t^e_2) = \begin{cases} \check{v}_\omega(t_0, t^e_2, v_\alpha(t_0)) & \text{if } \omega \neq \alpha_1 \text{ and } \omega \neq \alpha_2 \\ \check{v}_\omega(t^e_1, t^e_2, \hat{v}_\omega(t^s_1, t^e_1, \check{v}_\omega(t_0, t^s_1, v_\alpha(t_0)))) & \text{if } \omega = \alpha_1 \\ \hat{v}_\omega(t^s_2, t^e_2, \check{v}_\omega(t_0, t^s_2, v_\omega(t_0))) & \text{if } \omega = \alpha_2 \end{cases} \tag{4.4}$$

Here, t_0 is the reference point of time, i.e. the point of time before the execution of the first activity α_1. Note that the end time of an activity t^e_i is not necessary equal to the start time of the following activity t^s_{i+1}. The time period between two activities is the time needed to travel between the two activities. All state values are decreasing when the agent is travelling. This leads to additional percentage-of-time discomfort. Equation 4.4 shows the case for two dimensions. The equations stated are not valid, if α_1 and α_2 is the same activity. An extension to more dimensions is possible following the recursive rules defined in Equation 4.1 and Equation 4.2.

The duration discomfort for an activity α_i is defined in a simpler way since the duration state value is not dynamic. Therefore, no (recursive) calculations are needed to derive the duration discomfort. Following the definition of activity start times t^s and activity end times t^e from above,

the duration discomfort with a planning horizon n is defined as:

$$D_{dur}(\alpha, t^s, t^e) = \sum_{i=1}^{n} (T_{dur}^{\alpha_i} - (t_i^e - t_i^s))^2 * \gamma_{\alpha_i} \tag{4.5}$$

The duration discomfort of the execution of a single activity is dependent on a single value, namely the duration $(t_i^e - t_i^s)$ of activity α_i. The discomfort is the normalized quadratic difference with the corresponding duration target.

The decision procedure is the following: Whenever a decision about the next activity of an agent has to be made, all possible combinations of next activities are computed. The number of planned activities, the planning horizon, is a parameter of the simulation. The next step is the calculation of the activity duration minimizing the discomfort value at the end of the planning horizon, i.e. the following optimization problem is solved.

$$\begin{aligned} &\min_{\alpha, t^e} D_{pot}(\alpha, t^s, t^e) + D_{dur}(\alpha, t^s, t^e) \\ s.t. \quad & t^e - t^s \in \mathbb{R}_+^n \\ & t_i^e + ttime(\alpha_i, \alpha_{i+1}) = t_{i+1}^s \quad \forall i \in \{1, .., n-1\} \end{aligned} \tag{4.6}$$

Here, $ttime(\alpha_i, \alpha_{i+1})$ is the time needed to travel from activity α_i to α_{i+1}.

Note that, even with a fixed sequence of activities, finding the optimal duration is a non-linear, non-convex, multi-dimensional (when the planning horizon is larger than 1) optimization problem. This combination of attributes makes it very hard to find an analytic solution to the optimization problem which could be implemented in the simulation. Thus, a numerical solver to the problem is proposed and described in the next section.

4.3 Numerical Solver

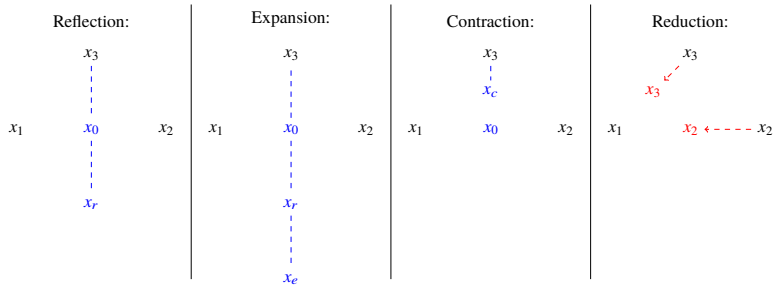
The main objectives of an algorithm which solves the full discomfort minimizing problem for a single agent and a fixed combination of activities, are fast runtime and good scalability for the planning horizon. These goals ensure that the overall simulation runtime does not increase to an insufficient level, when the planning horizon is extended.

The discussed Problem 4.6 is a multidimensional and non-linear opti-

mization problem. The description in the last section has shown that an analytic solution can not be provided and a numerical solver has to be implemented instead. The solver implemented in C-TAP is a direct search method. Direct search methods are advantageous for our purpose, because they do not need a derivative. They rely on evaluations of the optimized function which is the full discomfort function for the case of C-TAP.

The direct search method implemented here is the *Nelder-Mead* algorithm, sometimes also called Downhill-Simplex algorithm (Nelder and Mead, 1965). The structure of the algorithm is simple and is shown in Algorithm 4.1. The input for a n-dimensional problem contains three parts. First, $n + 1$ points in the solution space are needed. In our case a point is a vector of possible durations for the considered activity sequence. Second, the optimized function f has to be provided which is the full discomfort evaluation function. Finally, the algorithm stops when the function values of the considered points do not differ more than a given Δ. This is one of the possible ways to abort the main loop of the algorithm.

Figure 4.2: Illustration of the Nelder-Mead Algorithm



One iteration of the Nelder-Mead algorithm for two dimensions is illustrated in Figure 4.2. After sorting the $n + 1$ points by their function value (assume for the figure: $f(x_1) < f(x_2) < f(x_3)$) the centroid x_0 of the first n points is computed. After that, points of reflection ($\rightarrow x_r$), expansion ($\rightarrow x_e$) and contraction ($\rightarrow x_c$) are calculated. If none of the new points improves the current set by replacing x_{n+1}, reduction is performed (all points are moved towards the best point x_1). In summary, one iteration

Algorithm 4.1: Nelder-Mead Algorithm

```
1:  while max (x_i) − min (x_i) > Δ do
2:      ORDER vertices s.t. f(x_1) ≤ f(x_2) ≤ ··· ≤ f(x_{n+1})
3:      calculate the centroid x_0 of x_1, x_2 ··· x_n
4:      REFLECTION: x_r = x_0 + α(x_0 − x_{n+1})
5:      if f(x_1) ≤ f(x_r) < f(x_n) then
6:          x_{n+1} ← x_r
7:      else
8:          EXPANSION:
9:          if f(x_r) < f(x_1) then
10:             x_e ← x_0 + γ(x_0 − x_{n+1})
11:             if f(x_e) < f(x_r) then
12:                 x_{n+1} ← x_e
13:             else
14:                 x_{n+1} ← x_r
15:             end if
16:         else
17:             CONTRACTION: x_c ← x_0 + ρ(x_0 − x_{n+1})
18:             if f(x_c) < f(x_{n+1}) then
19:                 x_{n+1} ← x_c
20:             else
21:                 REDUCTION:
22:                 for all x_i, with i ∈ {2, ··· , n + 1} do
23:                     x_i ← σ(x_i − x_1)
24:                 end for
25:             end if
26:         end if
27:     end if
28: end while
```

computes different points looking for improvement of the maintained set of points.

The algorithm is driven by four parameters (α, γ, ρ and σ). The values $\alpha = 1$, $\gamma = 2$, $\rho = -\frac{1}{2}$ and $\sigma = \frac{1}{2}$ are known to be efficient for many problem descriptions (Lagarias *et al.*, 1998) and are also chosen in our case.

The discomfort minimization problem has one characteristic that influ-

ences the implementation. The activity duration variables are bounded. The lower bound of all activity durations is 0. It has been shown that the most effective approach to handle boundaries is a transformation of the domain of the function (Le Floc'h, 2012). The domain of the discomfort function is transformed with a logarithmic function. The transformed domain is unbounded and can be used as input for the Nelder-Mead algorithm.

The value for Δ cannot be chosen arbitrary small, because it has a substantial impact on the run time. Furthermore, the algorithm might find a local optimum. Thus, the solver has to be restarted several times with different initial points in order to reduce the probability not to find the global minimum. Note that C-TAP is simulating human behavior. Therefore, an estimation algorithm like the one described here is sufficient, because it is expected to find fast a reasonable solution in acceptable run time.

4.4 Conclusion

The traditional implementation of C-TAP as presented in Chapter 3 has to be modified for the purpose of simulating long-distance travel demand. Several adaptions were introduced in this chapter. Evident modifications were the simulated time period, the considered activity types, destinations and modes. Furthermore, an extension of the planning horizon was introduced. The latter induced a change in the decision making framework. Full discomfort is now used as objective function. This change yields a need of a numerical solver which has been presented in the last part of this chapter.

The adaptions introduced provide a possibility to utilize continuous activity planning to simulate long-distance travel demand. However, data describing the real world is needed to initialize simulation. Sources describing long-distance travel are usually not covered by traditional household travel surveys. Thus, alternative data collection techniques have to be considered. Collection of long-distance travel data is discussed in Chapter 6. The main part of the simulation initialization, namely the generation of a synthetic population, is topic of Chapter 7. An application of C-TAP is given in 8.

Chapter 5

Destination and Mode Choice

The previous chapter presented the core ideas of a continuous activity planning of long-distance travel. The decision making as described so far is limited to activity selection and activity duration optimization. However, travel decisions usually involve further dimensions. Particularly, mode choice as well as destination choice are part of activity planning when considering long-distance travel.

This chapter presents an approach to involve joint decisions on mode and destination choice. More specific, the presented framework introduces a heuristic to make joint decisions on activity type, activity duration, travel mode and destination. For this purpose, the discomfort minimization approach as presented before is modified with respect to the new task.

This chapter is structured as follows. First, a literature review on mode and destination choice models is provided. Then, all variables that are relevant for the decision making are introduced and discussed. This is followed by a description of the extended activity planning model. Afterwards, a heuristic for the model is presented which improves the simulation runtime substantially. The work presented in this chapter is based on several conference papers, namely Janzen and Axhausen (2017b,a, 2018).

5.1 Previous Work

Destination choice as well as mode choice has been the focus of a vast amount of studies. Statistical models have been the main tools used to explain destination and mode choices. Nevertheless, most of the studies focus on destination choice in daily life (e.g. Southworth, 1981; He *et al.*, 2009; Zheng and Guo, 2008) or particular parts of daily life, e.g. shopping

(Adler and Ben-Akiva, 1976; Bekhor and Prashker, 2008)

Mode choice models have been studied for daily-life travel intensively, e.g. by Vovsha (1997). Special interest has been given to impact of habits (e.g. Klöckner and Matthies, 2004) or attitudes (e.g. Johansson *et al.*, 2006). Alternative models like the regret minimization approach (Chorus *et al.*, 2008) have been taken into account as well. Discrete choice models in the scope of mode choice for long-distance travel have been investigated in some studies. For example, De Lapparent *et al.* (2013) implemented an MNL model for this purpose and drew conclusions about the value of travel time savings in Europe. Bhat (1995) used an extreme value model to explain mode choice for long-distance travel in the USA.

Several frameworks have been implemented in order to model mode choice in agent-based transport simulations. First, simple mode choice model using the random utility framework has been applied (Miller *et al.*, 2005). Second, a score for mode choice has been used in an agent-based simulation where scoring of activity plans drives the decisions of agents (Rieser *et al.*, 2009). Recently, a traditional discrete choice model was paired with an agent-based simulation to improve mode choice within the simulation (Hörl *et al.*, 2019). Nevertheless, all of these frameworks are applied to simulations of short-distance travel. Furthermore, destination choices in daily-life have been modelled in agent-based simulations as well. For instance, locations of secondary activities like shopping or leisure activities have been chosen with regards to the best fit into the rest of the schedule (Horni *et al.*, 2012; Horni, 2013).

Destination choice is a task that is more complex than mode choice, because the number of alternatives is usually substantially higher. Therefore, destination choice set generation has been studied intensively. The focus in choice set generation has been on vacation destination (Crompton, 1992; Crompton and Ankomah, 1993; Karl *et al.*, 2015). Besides choice set generation, modelling vacation destination choice is a problem tackled recently as well (Van Nostrand *et al.*, 2013; Bhat *et al.*, 2016; LaMondia *et al.*, 2010). However, these studies are limited to vacation destinations and long-distance destination choice has not yet been implemented in an agent-based simulation.

Finally, joint mode and destination choice has been focus of research. The apparent solution to this problem, a sequence of discrete choice models has been the approach widely used, e.g. in a travel demand model for the

San Francisco area (Jonnalagadda *et al.*, 2001). The sequential approach has been recently employed as well in case of long-distance travel, e.g. in the United States (Outwater *et al.*, 2015b). A joint MNL model for destination and mode choice has been implemented for the Chicago area (Anas, 1981). Nevertheless, a model for joint destination and mode choice has not been widely used in agent-based simulations, in particular not in agent-based simulations of long-distance travel demand. The decision making model described in this chapter is aiming to close this gap.

5.2 Decision Variables

The activity planning as described in the previous chapters does not imply a destination or mode choice. The destination has to be chosen simultaneously with the activity since it has direct impact on the discomfort of the agents. For example, the optimal activity duration of an activity can differ for different locations of the activities performed and the travel mode used. Destination and mode choice are dependent on various parameters, i.e. destination and mode specific parameters as well as individual agent parameters. Research has been studying intensively these parameters and their influence (Karl *et al.*, 2015; Crompton and Ankomah, 1993). The parameters involved in the decision making of C-TAP are presented in the following.

Activity planning which includes destination choice needs to utilize destination specific parameters. The main destination specific parameter is the location. A location in C-TAP can be defined either by coordinates or a node in a network. Locations are used to calculate travel distances in the corresponding network. It is distinguished between location and destination to add more flexibility to the model. A location has no semantics and is solely used to describe travel distances and, therefore, indirectly travel durations. On the other side, a destination has more information than a physical location. This information includes costs of the destination visit which are dependent on the duration of the activity performed at this destination.

Each destination has also a calibrated attractiveness value for each activity type. Attractiveness describes the objective part of agents' gravitation to certain destinations. For example, beach destinations attract many vacations

while they are not very attractive for business travellers. A subjective part of attractiveness, the perception, is implemented as well (see Section 5.2.2). Furthermore, an overnight stay at a destination has a certain monetary cost (mainly for hotel). This cost usually applies to vacation trips since travellers mostly do not pay for themselves for business trips. Other parameters have been studied, but are not included here. For example, the need of a visa for specific countries or the question whether the destination is domestic or international may influence destination choices.

A second type of decision parameters are mode specific, i.e. they depend on the mode considered in the decision making. There are two important decision drivers in mode choice. First, the price of travel has to be considered. More specific, the monetary costs of getting to a certain location have to be known. These costs are modelled as a mode specific function of distance. This function has not to be linear, as it is the case for air travel. Second, the crucial parameter is the travel speed. Similar as above, the travel time is a mode specific function of the distance which has not to be linear. Further parameters like comfort are usually involved in mode choice models in order to get a detailed model. However, mode choice in long-distance travel is relatively simple and is mainly driven by the distance travelled. Long distances are covered by airplanes while shorter distances are travelled by car or sometimes by rail. This mode choice is covered by the two mode parameters described above. Therefore, there is no need for a more parameterized mode definition within C-TAP.

Besides destination and mode parameters, individual parameters play a substantial role in activity planning. First of all, all decisions made in C-TAP focus on discomfort minimization as described in Chapter 4. Thus, the targets of each agent are the core of each decision in the activity planning. Percentage-of-time and duration targets for each activity of an agent are implemented and used to calculate discomfort values. Full discomfort is employed to evaluate and rank the choices of each agent. However, further agent parameters potentially change the choice sets. Mode accessibility effects mode choice as well as indirectly destination choice. Mode accessibility covers two different aspects. On the one hand, car availability has direct impact on the mode choice set. On the other hand, access time to airports or train stations impacts mode choice indirectly. Parameters, which are not included so far, are socio-demographics like age or education. Income is modelled indirectly via budgets as shown in

Section 5.2.1. Additionally, a certain destination loyalty was observed in recent studies (Niininen *et al.*, 2004; Oppermann, 2000; Oom do Valle *et al.*, 2006). Loyalty is not taken into account in C-TAP since loyalty is observed during the course of several years. However, C-TAP is designed to simulate a single year of long-distance travel demand.

The destination choice model of C-TAP includes all the parameters mentioned above. Before the heuristic implementation of the model is shown later in this chapter, some of the parameters and their implementations are described in more detail in the following. An application of activity planning with all parameters described above is shown later in the application (see Chapter 8).

5.2.1 Budget Constraints

The introduction of budgets to C-TAP is an extension to the traditional need-based and target-based models. Budgets are essential to decisions of persons, because they limit their choice set. For example, the available amount of money is one of the main drivers of vacation destination choice. Therefore, it is valuable to add budgets to an agent-based simulation of long-distance travel demand. Two different types of budgets are taken into account within the C-TAP activity planning, when a non-business activity is considered. First, monetary budgets for each agent are used to implicitly model income effects on the decision. Second, time budgets can be added to the budget set.

A monetary budget can be used to cover the costs of long-distance travel, i.e. the travel cost depending on the travel mode and travel distance as well as as the activity costs depending on activity type, activity destination and activity duration. The budget is related to the income of the virtual agents modelled in the simulation. It can be either a monthly growing value modelling monthly savings of the agent or a static budget defined prior to the simulation start for the whole year simulated.

Time budgets are considered during the decision making and set boundaries for the choice sets. The most prominent instance of a time budget is the number of vacation days granted by the employers. The vacation days define a budget that has to be used each time an agent is performing a non-business activity during a day which is neither a weekend nor a holiday. The implementation of a vacation day budget ensures that most

of the long-distance trips are performed on weekends which is reasonable. Vacation day budgets have less impact on long vacations since they cover both weekdays and weekends. On the other hand, business activities are not affected by vacation time budgets at all. Therefore, they can be performed at any time without regard to time budgets. However, there is a concept to push business trips away from weekends which is the seasonal attractiveness. The idea of seasonal influences is presented in Subsection 5.2.3.

Budgets have an impact on activity planning in two different ways. First, a budget defines a hard constraint, i.e. an agent can not exceed the given budget. For example, it is not possible to spend more money than it is available in the monetary budget. Second, each budget is also used as a soft constraint. In other words, consuming a part of a given budget results in disutility for the corresponding agent. Disutility in C-TAP is measured with the discomfort value. Therefore, a budget discomfort is added to the percentage-of-time discomfort and the duration discomfort that were presented in Chapter 3. Budget discomforts are weighted by simulation time left if the budget is static since the consumption of a large part of a budget is more crucial at the beginning of the simulation time. For example, using all vacation days in the beginning of the year simulated gives a higher discomfort than using them towards the end of the year.

5.2.2 Individual Preferences

Activity planning models describe human behavior which is not deterministic. Heterogeneity has to be added to the activity planning and, in particular, to the destination choice of the virtual population. Therefore, the choices are not solely driven by socio-demographics or externalities. A concept that is related to the idea of mental maps (Narayana and Markin, 1975; Swiderski, 1983; Hannes *et al.*, 2008; Ordóñez Medina, 2015) is included in C-TAP. Mental maps are used to describe the limited view on the set of destinations that a person usually has. This set of destinations is generally not the full set since a person in reality never has full information as it is assumed in simple models. Two different ideas are implemented in the mental maps of the population in C-TAP. First, destination awareness indicates whether an agent is aware of the destination. Second, destination perception describes the perceived attractiveness of the destination.

Awareness has been shown to be an important part in the decision process

since most of the people are actually not aware of all destinations that are accessible (Karl *et al.*, 2015; Seddighi and Theocharous, 2002). Destination awareness is a binary function that reveals for each pair of agents and destinations whether the agent is aware of the corresponding destination. If the agent is not aware of a destination, this destination is not considered in his activity planning. Thus, awareness causes an agent to consider a subset of the set of destinations during his decision making.

Individual perception is used to account for variations in subjective perceptions of destinations. The perception of a destination is an individual function reducing or increasing the attractiveness of each destination. The perceived attractiveness of a single destination may differ substantially among the set of agents. Therefore, agents with exactly the same objective parameters can make different destination decisions since they perceive destinations in different ways. Hence, additional variation is added to the simulation.

Efficient implementation of both, awareness and perception, is not trivial. Storing these two values for each pair of agents and destinations is extremely inefficient since the set of agents as well as the set of destinations are potentially very large. It is also not possible to store a static function returning these two values since they do not follow specific rules that can be translated to a function. Instead a single random number is attached to each agent. Two hash functions map the random number in combination with a destination to a binary awareness value and an individual perception value. This technique ensures that awareness and perception are constant for each agent-destination pair throughout a simulation run without storing these values for each pair.

5.2.3 Seasonal Influence

The parameters presented in this section follow different types of dynamics. Some of them are purely static, e.g. destination awareness is fixed for each agent-destination pair throughout the whole simulation. Many parameters are dynamic, but defined by the agent history, e.g. percentage-of-time state values are determined by the activities performed. However, a third class of dynamics is needed, namely parameters that are externally dynamic. These parameters are part of the initialization of the C-TAP simulation.

Mode specific parameters that are dynamic are travel times for car travel

and travel costs for air travel. Both values increase at certain times of the year, in particular at the beginning and end of school holidays. Furthermore, destination attractiveness is a dynamic parameter as well. The day of week has an influence on the attractiveness of business activities since weekends are unlikely to attract business travellers. On the other hand, the day of the year has an influence on the attractiveness of vacations since some destinations, like the Mediterranean Sea, does not attract many vacation travellers during winter time. Therefore, destination attractiveness can be defined as a product of several functions combining weekday attractiveness and season-of-the-year attractiveness to a single value.

All seasonal parameters are potentially part of the calibration process which is described in detail in Section 9.2. An application is presented in Chapter 8.

5.3 Modelling Destination and Mode Choice

The attributes that are implemented in C-TAP have been described above. The question remains how this attributes are taken into account in the target-based activity planning as it was described in Section 4.2. This question will be tackled in the remainder of this section.

Before introducing the optimization problem some terms that are used for the rest of the chapter are given in the following. The starting condition of a decision making problem is a pair (λ_0, t_0) which is defined by the current location λ_0 of an agent and the point of time t_0 before the decision is made. A potential solution of the decision making problem discussed here is called alternative. In the following, an alternative (α, λ, μ) consists of an activity α, a destination λ and a mode μ. An optimized alternative includes optimal end time t^e of the activity execution. In this case, optimality is equated with minimal full discomfort. A feasible optimized alternative is an optimized alternative that can be performed by the agent. Feasibility includes that an agent has α in his activity set, μ is available to it and λ can be reached with mode μ before t^e. Given a planning horizon of $n > 1$, an alternative becomes a triple of n-dimensional vectors α, λ and μ. The n-dimensional alternative describes a sequence of activities, destinations and travel modes.

In the following the two additions are described that mutate the decision

making task as presented in Section 4.2, namely the destination pull factor and the budget discomfort. Afterwards, the new optimization problem is described and a heuristic solution approach is given.

5.3.1 Pull Factor

Some of the variables described above reduce the probability to visit a specific destination λ. In order to model this probability, a pull factor ϕ_λ is introduced. The pull factor combines the attractiveness of a location s_λ and the agents perception for a location $\varepsilon_\lambda \in [0.5, 1.5]$. As described in Section 5.2.3, the attractiveness is a function of time in order to represent seasonal influences. Hence, attractiveness is given by a function $s_\lambda(t) \rightarrow [0, 1]$. The range of s_λ is given by an interval between 0 (no attractiveness) and 1 (maximal attractiveness). The perception can take a value above 1, i.e. the perceived attractiveness of a location might be higher than the actual attractiveness. The perception value is not a seasonal function, but is a fixed value for each pair of agent and destination throughout a simulation run.

The attributes $s_\lambda(t)$ and ε_λ need to be included in the decision making of C-TAP agents. Assuming an agent performs an activity at λ with starting time t^s and end time t^e, a pull factor is defined:

$$\phi_\lambda(t^s, t^e) = \varepsilon_\lambda * \int_{t^s}^{t^e} s_\lambda(u) \frac{\mathrm{d}u}{(t^e - t^s)}. \tag{5.1}$$

The integral over the activity execution time gives an average of the attractiveness weighted by the individual perception of the corresponding agent. The pull factor summarizes with a single value the destination attractiveness with regards to individual perception as well as the seasonal influence. A higher pull factor has to lead to a higher probability of the corresponding agent to perform the activity at the time and location related to the high pull factor. A higher probability for activity execution is achieved with lower discomfort as it was described in Section 4.2. Thus, the state values of the considered activities have to be higher when considering alternatives with high pull factors. The state value increasing function is

modified by multiplying the pull factor to the exponent:

$$\begin{aligned}\hat{v}(t^s, t^e, v_0) &= 1 + (v_0 - 1)e^{-\tau \cdot \phi_\lambda(t^s, t^e) \cdot (t^e - t^s)} \\ &= 1 + (v_0 - 1)e^{-\tau \cdot \varepsilon_\lambda \cdot \int_{t^s}^{t^e} s_\lambda(u) \mathrm{d}u}\end{aligned} \tag{5.2}$$

A higher value ϕ_λ leads to an increase of the slope of $\hat{v}$. Hence, the state value raises faster for the considered activity. As a consequence, the activity is more likely to be executed at locations with high ϕ_λ-values. A pull factor close to zero (e.g. due to a seasonal effect) will prevent the corresponding state value from rising. Therefore, an activity execution at these locations would not contribute to a discomfort reduction. Consequently, the activity will not take place there.

The pull factor merges two location attributes to describe attractiveness of a specific destination. However, it is worth noting that these two attributes differ substantially. The destination attractiveness function s_λ is unique for each activity while the perception value ε_λ is unique by agent. The pull factor combines these two attributes from two different sources.

5.3.2 Budget Discomfort

Budgets are implemented in C-TAP and were introduced in Section 5.2.1. Monetary budgets as well as time related budgets set hard limits for the solution space and reduce the set of valid alternatives for an agent that has to make a decision on his next activity.

However, staying within budget boundaries is not sufficient when considering monetary budgets. An agent has to try to minimize his costs. In other words, it has to save as much as possible of its budget. This is certainly important when comparing similar alternatives that differ by their costs. In this case, an agent should always prefer the cheaper alternative. This behavior can not be achieved with a hard constraint. Hence, an additional discomfort type is added to the simulation, namely the budget discomfort.

Assume that the costs of travelling between two destinations λ_i and λ_{i+1} ($i \geq 1$) costs $c_M(\mu_i, \lambda_i, \lambda_{i+1}, t_i^e)$ when using mode μ_i at point of time t_i^e. Furthermore, assume that the costs of a non-business activity α_i with start time t_i^s and end time t_i^e are given by $c_D(\lambda_i, t_i^s, t_i^e)$ when performing the activity at destination λ_i. Time consuming the time budget (e.g. vacation days) in the time interval t_x to t_y is given by $c_V(t_x, t_y)$. Thus, the budget

discomfort for a feasible alternative (α, λ, μ) and activity end time t^e with time horizon n is defined as

$$D_{bud}(\lambda, \mu, t_e) = \frac{(\sum_{i=1}^{n} c_D(\lambda_i, t_i^s, t_i^e) + c_M(\mu_i, \lambda_{i-1}, \lambda_i))^2}{B_C^2} + \frac{c_V(t_0, t_n^e)^2}{B_V^2}. \quad (5.3)$$

As before the starting condition is given by the pair (λ_0, t_0). B_C is the monetary budget and B_V is the time related budget of the corresponding agent.

The budget discomfort equals zero when an alternative does not involve any monetary costs or vacation days. The discomfort grows with an increase of costs or vacation days. Thus, agents are pushed towards budget saving alternatives. Nevertheless, budget discomfort is not intended to be a decision driver, but a tie breaker when it comes to compare two similar alternatives. Otherwise, agents would avoid any vacation in favor of business trips since these do not reduce any budget. Thus, the budget discomfort has a lower weight within the discomfort minimization. The weighted budget discomfort ensures that agents choose the cheaper option, if there are two similar locations.

5.3.3 Optimization Problem Definition

The decision making problem as shown in Section 4.2.1 has to be modified when taking into account the extensions introduced in this chapter. On the one hand, the decision to be optimized is expanded by two dimensions, mode and destination. On the other hand, additional constraints as budgets and additional parameters as awareness are included.

Consequently, the discomfort minimizing problem can be expressed as

$$
\begin{aligned}
\min_{\alpha,\lambda,\mu,t_e} \quad & D_{pot}(\alpha,\lambda,t^s,t^e) + D_{dur}(\alpha,t^s,t^e) + \gamma_b D_{bud}(\lambda,\mu,t^e) \\
s.t. \; & t^e - t^s \in \mathbb{R}^n_+ \\
& \sum_{i=1}^{n} c_D(\lambda_i, t_i^e - t_i^s) + c_M(\mu_i, \lambda_{i-1}, \lambda_i) \leq B_C \\
& c_V(t_0, t_n^e) \leq B_V \\
& t_i^e + ttime(\alpha_i, \alpha_{i+1}) = t_{i+1}^s \quad \forall i \in \{1,..,n-1\}
\end{aligned}
\tag{5.4}
$$

where the optimized variable is a feasible alternative with a sequence of activities α, destinations λ, travel modes μ and activity end times t^e. As before, the starting condition is given by (λ_0, t_0). The percentage-of-time discomfort is dependent on the destinations of the performed activities since they impact the discomfort via pull factor ϕ. Additionally, the monetary budget B_C and the time budget B_V limit the choice set with respect to the corresponding costs. The budget discomfort D_{bud} is weighted by a factor $\gamma_b < 1$ to ensure that the budget is not the main driver for the decisions made. Further constraints that are not noted in Problem 5.4 include destination awareness and mode accessibility. These constraints are excluded since only feasible alternatives are considered.

In C-TAP, Problem 5.4 has to be solved every time an agent becomes idle and has to decide on his next activity. Activity type, activity destination, activity duration and travel mode are optimized simultaneously. The computation of the solution has to be fast, because the number of agents is sizeable and a reasonable time simulated is a year. Hence, the number of optimal decisions that has to be calculated can be enormous. In addition, the planning horizon used in C-TAP includes usually more than a single activity in advance. Each additional activity in the planning horizon adds multiple dimensions to the optimization problem. Therefore, a heuristic approach is needed to solve the minimization problem in reasonable time. Techniques for an efficient implementation of the decision making process are presented in the following section.

5.4 Efficient Implementation

An efficient implementation of the decision making is crucial for a reasonable simulation run time and therefore for the applicability of C-TAP. Due to its efficiency and flexibility the core of the optimization process in the activity planning remains the Nelder-Mead algorithm which was introduced in Section 4.3. It has been shown that the Nelder-Mead algorithm computes an estimate of the optimal duration for a given activity or a given sequence of activities. The extensions of destination and mode choice do not mutate the duration optimization problem since all combinations of modes and destinations need to be enumerated. The duration needs to be optimized for all feasible alternatives. The number of alternatives can be potentially very large since all activities, destinations and modes have to be enumerated. Therefore, the main idea to speed up the decision making is the reduction of the number of alternatives that are considered. Two sequential procedures are implemented to reduce the number of alternatives considered. First, a heuristic is used to analyze the characteristics of each alternative and subsequently to decide which alternatives are likely to be optimal. Second, the optimal discomfort, i.e. the discomfort with optimal durations, is estimated for each of the remaining alternatives. Only those alternatives that have low estimated discomfort are kept and are considered further. Both ideas are discussed in detail in the following.

5.4.1 Selection Heuristic

The mathematical problem formulated above can not be solved optimally in reasonable time. Note that the set of destinations is discrete, unordered and potentially enormous. Though, the problem has to be solved every time an agent has to make an activity decision. Thus, it is necessary to have a solver that provides a result fast. Therefore, an heuristic approach is implemented. The main idea is a reduction of the discrete solution space. A smaller set of feasible choices simplifies the optimization problem.

The activity planning process of an agent starts when the agent finishes an activity and turns idle. A decision on the next activity has to be made at this point. The decision making starts by generating the full set of all combinations of activities, modes and destinations. A triple of activity, mode and destination is called alternative. For larger planning horizons an

alternative is a sequence of triples. As mentioned above, this set is usually too large to find optimal durations for each alternative in reasonable time.

Several heuristic steps are implemented to reduce the set of alternatives and therefore speed up the activity planning. First, all hard constraints for modes and destinations are taken into account. Particularly, alternatives with destinations which the agent is not aware of are removed as well as the alternatives with car travel if the agent has no car available. Second, budgets and their implications are considered. In other words, it is checked whether it is likely that an alternative is feasible with regards to the budget constraints after duration optimization. Since the optimal duration is not known one needs an estimation of it to evaluate the budget constraints. The duration target of the corresponding activity can be taken as an estimate of the optimal duration since an agent tries to meet the duration target in order to reduce the duration discomfort. Given the target one can estimate the costs for the considered alternative by assuming the target duration will be met. If the costs exceed the budget substantially, the alternative is not further considered. This approach can be applied to both, monetary budgets and time related budgets.

Furthermore, the pull factor ϕ_λ is a good indicator whether an agent is likely to visit a specific destination λ. Considering a single activity, there is no need to consider all destinations in the activity planning. An agent will usually not perform an activity with a low pull factor if there is an alternative with a higher pull factor. Thus, all alternatives that have a pull factor below a certain threshold (e.g. 0.5) are removed. This approach can theoretically lead to an empty choice set, because all alternatives have low pull factor. In this particular case, fallback is triggered, i.e. no alternative is removed.

Finally, one can exclude specific modes from the choice set. Neither are planes usually used for short distances nor are trains utilized for intercontinental travel. Thus, the following procedure is implemented to identify the modes that are likely not to be used. For all alternatives with the same destination the fastest mode to get to this destination is identified. This is the reference mode for this destination. Then, all alternatives are identified which have the same destination and a mode that is slower than a certain factor, e.g. 1.2, of the reference mode. These alternatives are removed from the choice set. This procedure will remove all modes but the fastest for most destinations from the choice set, because the range of

distances where several modes have similar travel times is usually small.

The sequential listing of a step-by-step reduction of the choice set suggests that the reduction is implemented in a particular order. The constraints for the set of alternatives described so far are taken into account while generating the choice set. This approach saves memory and simulation runtime. Applying all heuristic steps leads to a reduction of the number of considered alternatives. However, the remaining set may still have substantial size. Therefore, further alternatives have to be removed to assure an efficient computation of the optimal alternative.

5.4.2 Promising Alternatives

Implementation of the selection heuristic described above leads to a substantial reduction of feasible alternatives considered in the activity planning. Nevertheless, the set of alternatives can still be too large to run the Nelder-Mead algorithm for optimizing all alternatives. Therefore, it can be beneficial to further reduce the set of alternatives before running the optimization. The main idea is to estimate the discomfort of the optimized alternatives before the optimization.

The final activity planning decision of an agent is based on the discomfort of the optimized alternatives. The discomfort is not known before the optimal activity duration is calculated. However, the discomfort can be estimated by calculating the discomfort of the alternative with a duration that meets the corresponding duration target. There is no evidence that the estimated discomfort is close to the optimized discomfort. Nonetheless, one of the objectives of the activity planning is to reduce the duration discomfort which is achieved by execution of an activity for a duration that is close to the duration target. Thus, it can be assumed that the estimated discomfort is a good approximation of the actual discomfort of this alternative.

Calculating a single discomfort value for each alternative can be done quickly as it is an application of few mathematical operations (see Equation 5.4). Having done that, one can sort all alternatives by their estimated discomfort. The order is expected to be the same order as after optimizing the alternatives. Hence, the alternatives with the lowest estimated discomfort are kept, the so called promising alternatives. Calculation of the optimal duration can be limited to these promising alternatives. Activity planning can be speed up substantially in dependence on the number of promising

Table 5.1: Impact of number of alternatives on optimality and run time

Number of alternatives	Best alternative	Share best alternatives	Optimality gap (avg.)	Simulation run time
1	504,283	94.7%	13.5%	28min
2	509,886	95.7%	13.1%	49min
3	516,666	97.0%	14.3%	1h 08min
4	519,752	97.6%	16.1%	1h 25min
5	523,177	98.2%	18.9%	1h 42min
6	526,235	98.8%	15.4%	2h 19min
All(7+)	532,666	100.0%	0.0%	3h 55min

alternatives kept. Since activities may have substantial differences in their characteristics the number of promising alternatives is defined per activity type. This approach makes sure that each activity type is represented in the set of promising alternatives.

A low number of promising alternatives kept can improve the run time of C-TAP while the impact on the decision making is not clear without further analysis. The trade-off of simulation run time and optimality gap can not be quantified analytically. However, one can study the trade-off in a test-case. For this purpose, a simulation with a test-case was set up. The simulation includes 10,000 agents, 10 different destinations and 3 modes. No awareness was implemented and all other agent attributes and destination attributes were randomly drawn. The modes are air travel, car and train.

Table 5.1 shows the simulation results of this test-case. In total, the agents had to make 532,666 decisions. As one can see in the last row of the table, the optimal solution is always found and the run time of the simulation is almost 4 hours if all alternatives are kept. Reducing the set of promising alternatives to one alternative per activity reduces the runtime by factor eight which is substantial. The optimality gap is between 13 and 18 percent if not all alternatives are considered. However, this applies only to a very small share of decisions. Considering only one promising alternative still gives the optimal solution in almost 95 percent of all decisions. This fact justifies the promising alternatives approach.

The number of promising alternatives can be controlled via parameter by the simulation user. Considering the results presented here, a low number of promising alternatives is suggested, especially when dealing with large scenarios.

5.4.3 Illustration

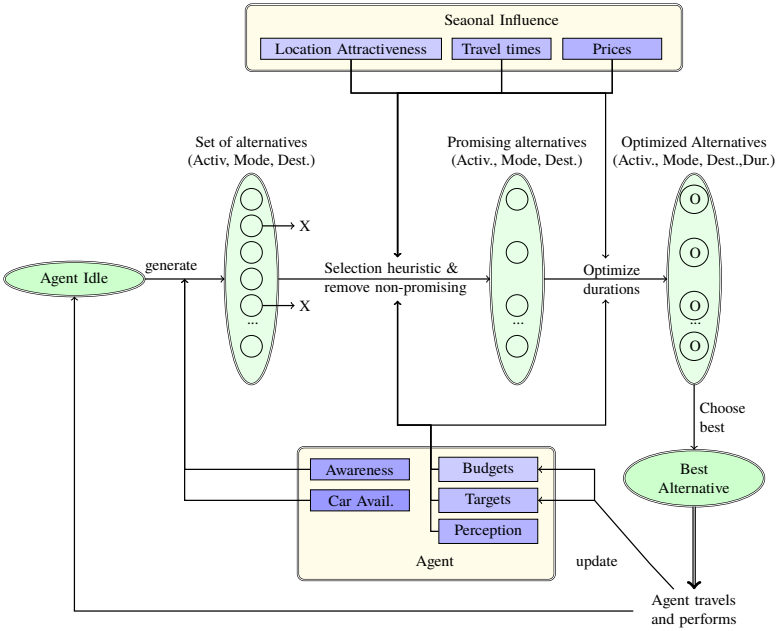
The activity planning including destination and mode choice was presented in this chapter focusing on several aspects of the framework. A summary of the activity planning is given in the following and is sketched in Figure 5.1. The activity planning is triggered every time an agent finishes an activity and becomes idle. An idle agent needs to decide on his next activity. In a first step, a set of alternatives for his next action is created. Potentially, an alternative is a sequence of activities, destinations and modes. Afterwards, the selection heuristic described in Subsection 5.4.1 is applied. Alternatives, which can not be afforded by the agent, alternatives with a low pull factor ϕ_{λ} and alternatives with inefficient travel modes are discarded. Subsequently, the non-promising alternatives are removed as described in Subsection 5.4.2. Then, optimal duration is computed for each alternative in the remaining set (see Equation (5.4)). A detailed description of the discomfort minimization approach was given in Subsection 5.3.3. Finally, the agent performs the first activity of the best alternative from the optimized set. After executing the chosen activity targets and budgets are updated.

5.5 Discussion and Conclusion

A destination and mode choice approach has been presented that can be used within a continuous target-based microsimulation for long-distance travel demand. Three stages were implemented. First, a heuristic reduces the number of considered destinations. Second, non-promising alternatives are removed from the alternative set considered. Last, the discomfort minimizing solution is computed among the remaining alternatives.

A major concern of this approach is data availability, and thus, calibration of the simulation. Microsimulations as C-TAP need individual data for calibration and validation of the software. In case of long-distance travel, these data sources are very rare and usually have small sample size. Therefore, alternative data sources (e.g. GPS or GSM data) have to be taken

Figure 5.1: Sketch of an agents decision process



into account in order to increase the value of the agent-based simulation presented in this work. An alternative data source and its potential is discussed in the following chapter.

Due to the heuristic steps implemented in the activity planning, the number of possible destinations in C-TAP is higher than usual in activity-based simulations. Nevertheless, handling thousands of destinations (as it is the case in real world) is still too resource consuming. Additionally, several C-TAP destinations can share the same location which further adds to the complexity. However, the approach is applicable to the top-level of a hierarchical destination choice problem, e.g. choice of a country or region before the choice of the actual destination. An application of C-TAP including mode and destination choice is presented in Chapter 8.

Chapter 6

Data Collection

In order to develop tools which are able to provide reliable predictions, one needs data sources that describe the current state of long-distance travel demand. This applies also to C-TAP. Initialization, calibration and validation of the simulation requires reliable input data.

Data collection methods in the field of travel demand research have been investigated in the past (Axhausen *et al.*, 2002a; Armoogum and Madre, 2002; Bonnel *et al.*, 2009; Zmud *et al.*, 2013; Richardson *et al.*, 1995; Arentze *et al.*, 2000; Draijer *et al.*, 2000). The most frequently used data sources are surveys. In case of long-distance travel, the number of available surveys is limited. The main sources are national travel surveys. Admittedly, all long-distance travel surveys face similar problems. Due to the high response burden surveys tend to have a low number of respondents. Furthermore, it is known that the number of journeys reported in such surveys is too low (Madre *et al.*, 2007; Armoogum and Madre, 2002). Both factors limit the explanatory power of the studies and leave the question of the quality of the results unanswered (Kuhnimhof and Last, 2009).

To overcome these limitations alternative data sources are needed. Mobile phone billing data can be used in order to obtain better estimates of long-distance travel demand. The advantage is the large number of individuals that can be tracked without having being asked to spend a lot of effort on a survey. Five months of mobile phone billing data covering one third of the total French population were analyzed for this dissertation. The data was provided by Orange™ France.

After reconstructing long-distance journeys from the data, the error reported by the French National travel survey is quantified. The main analysis is split in two parts. First, the number of persons that do not travel long distances at all is quantified. This analysis shows that there are more

non-travellers among survey respondents than among Orange™ customers. Second, the number of long-distance tours that are done by the mobile persons is quantified. It is shown that mobile Orange™ customers travel significantly more than survey respondents. Both results indicate that the number of tours was heavily under-reported in the survey. The aim of this chapter is to confirm the assumed under-estimation of long-distance tours and demonstrate that there is a need of alternative data sources. The utilization of this data for C-TAP is presented in the next chapter, where it is shown how mobile phone data is used to synthesize a population for long-distance travel demand.

This chapter is structured as follows. After a literature review the mobile phone data made available for this work is described in detail. This is followed by a description of the tour reconstruction methodology implemented. Afterwards, the results are compared with a traditional national travel survey. The findings presented in this chapter build on work published in Janzen *et al.* (2018).

6.1 Previous Work

Data collection has always been an important issue in the field of travel demand research. Different methods of data collection have been investigated in the past (Axhausen *et al.*, 2002a; Armoogum and Madre, 2002). The data sources used have mostly been various forms of surveys to suit the diverse requirements of the researchers (Dillman, 2000).

In the case of long-distance travel, the number of recent surveys is limited. For Europe, *Mobidrive* studies are available (Zimmermann *et al.*, 2001; Axhausen *et al.*, 2002b; Chalasani and Axhausen, 2004). Each of these studies encompasses a six-week period, which is usually not sufficient for a deep analysis of long-distance travel behavior. Other sources are national travel surveys like the French (Armoogum *et al.*, 2008), British (Department for Transport, 2016) or Austrian (BMVI, 2012) ones. An additional longitudinal perspective is provided by the INVERMO study from Germany (Chlond *et al.*, 2006). Several European studies have been combined for an analysis of long-distance travel demand in Europe (Frick and Grimm, 2014). A similar approach led to a nationwide model for the United States (Outwater *et al.*, 2015a,b; Bradley *et al.*, 2015).

An overview of available studies of annual long-distance travel rates can be found in Table 6.1 which reports the study area and year. Variations in the definition of long-distance travel are also given which include the distance-threshold used, the destinations included in the analysis and whether single-day tours were excluded from the set of long-distance journeys. Finally, the main indicator, the annual number of long-distance tours, reported in the studies are presented. The values that had to be extrapolated are marked. The studies included are: the California Statewide Household Travel Survey (CSHTS) (Bierce and Kurth, 2014; Cambridge Systematics Inc., 2013), an ifmo study (Frick and Grimm, 2014; Kuhnimhof *et al.*, 2014), the INVERMO project (Zumkeller *et al.*, 2005; Chlond *et al.*, 2006), the Knowledge Base for Intermodal Passenger Travel in Europe (KITE) (Frei *et al.*, 2010), the DATELINE study (Neumann, 2003), the French national travel survey (ENTD) (Armoogum *et al.*, 2008), the Microcensus Switzerland (MCS) (Swiss Federal Statistical Office (BFS) and Swiss Federal Office for Spatial Development (ARE), 2012) a Eurostat report (Weckström-Eno, 1999), Methods for European Surveys of Travel Behaviour (MEST) (Axhausen and Youssefzadeh, 1999), and US National Transportation Statistics (US NTS) (Bureau of Transportation Statistics, 2016). All of these studies surveyed 8–12 weeks of long-distance travel and estimated annual tour rates. A correction factor is incorporated in most of the tour rates. The ifmo study reports a higher value than the other studies due to several reasons. First, it is one of the most recent studies and it is known that the amount of long-distance journeys is growing. Second, it is combining several studies to get a full picture and, in particular, it estimates 5.0 everydays long-distance tours (e.g. commuting) which is more than in any other study.

Other long-distance travel studies have been performed with a special emphasis on tourism. Guidelines for tourism studies (Harris *et al.*, 1994) and preferred analysis methods (Crouch, 1994) have been presented in the past. Many tourism studies have been performed, including the Travel Market Switzerland study (Bieger and Lässer, 2008) and the Net Traveler Survey (Schonland and Williams, 1996). Almost all of them focus on tourism activities within a single country. A summary of international studies can be found in Lennon (2003) or the Eurostat database (Eurostat, 2016). However, the results of tourism surveys are limited due to the known issue of unobserved tourism (De Cantis *et al.*, 2015).

Table 6.1: Annual long-distance tour frequencies: Other studies (* based on own extrapolation)

Study	Year	Area	Destination	Long-dist. definition	Exclude single-day	Annual tours per capita
DATELINE (2003)	2001–02	Europe	international	100km	No	2.7
DATELINE (2003)	2001–02	France	international	100km	No	3.8
ENTD (2008)	2007–08	France	France	80km	No	5.1
MEST (1999)	1997–98	France	international	100km	No	*7.4
MCS (2012)	2010	Switzerland	international	100km	No	*7.8
MEST (1999)	1997–98	Europe	domestic	100km	No	*7.9
KITE (2010)	2008–09	Switzerland	international	100km	Yes	8.2
KITE (2010)	2008–09	Portugal	international	100km	Yes	8.2
CSHTS (2014; 2013)	2012	California	state-wide	50 miles	No	8.2
Eurostat (1999)	1999	France	international	100km	No	8.5
INVERMO (2006)	2001–03	Germany	international	100km	No	8.8
MEST (1999)	1997–98	Europe	international	100km	No	*8.9
KITE (2010)	2008–09	Czech Rep.	international	100km	Yes	9.0
US NTS (2016)	2001	USA	international	50 miles	No	*9.4
ifmo (2014; 2014)	2011	Germany	international	100km	No	15.9

Due to the high response burden that is usually associated with long-distance surveys (Axhausen *et al.*, 2015; Axhausen and Weis, 2010), it can be expected that the number of long-distance trips is often under-reported. This is due to non-responding frequent travellers as well as travellers claiming not to travel while answering other questions, so-called soft refusers (Madre *et al.*, 2007). Furthermore, there is a memory effect. Respondents tend to forget tours that happened some time before the survey (Smith and Wood, 1977; Bradburn *et al.*, 1987; Tourangeau, 1999). Additionally, the vehicle miles travelled are usually heavily underestimated as shown by Wolf *et al.* (2003). Consequently, there is a need for survey weighting and expanding (Bar-Gera *et al.*, 2009). Assumptions about under-reporting long-distance tour rates in surveys led researchers to introduce correction factors in several studies (Cambridge Systematics Inc., 2013; Armoogum *et al.*, 2008). In case of tourist surveys, a weight correcting for the response bias is essential (Leeworthy *et al.*, 2001). A correction factor is the main method currently available to account for under-reporting. A second idea to overcome this issue is the combination of several studies that are focusing on different aspects of long-distance travel as it has been done in Germany (Kuhnimhof and Last, 2009; Frick and Grimm, 2014) or California (Bierce and Kurth, 2014; Cambridge Systematics Inc., 2013). Assumptions about the inaccuracy of long-distance travel surveys are supported by evidence that two surveys of the same scope can suggest non-consistent travel behavior (Perdue and Botkin, 1988). A comparison of several studies led to the conclusion that models using household travel surveys need validation with independent data (Ashley *et al.*, 2009).

In order to estimate the level of under-reporting in surveys, one needs alternative data sources. Anda *et al.* (2017) show that big data has to be utilized nowadays in transport modelling. Two main big data sources are available for analyzing travel demand. Both use passive data collection. First, GPS data can be used to collect information about travel behavior (e.g. Montini *et al.*, 2014). Yet, the collection of GPS data is limited because cooperation of the respondents is needed, and smartphone GPS collection is battery-consuming and, thus, discouraging participation. Nevertheless, it has been already used to prove under-reporting of trips in travel surveys from New York (Jin *et al.*, 2014), California (Wolf *et al.*, 2003) or Kansas City (Bricka and Bhat, 2006). The latter study provided a regression model to confirm that socio-demographics have an impact on the probability not to

report trips in a survey. Second, mobile phone network operators produce billing information that provides an enormous amount of data. This has already been utilized in various fields (Blondel *et al.*, 2015) including transportation. One of the first applications was an analysis of travel demand induced by tourism (Ahas *et al.*, 2008a, 2007). GSM data has also been used to estimate OD-matrices (Friedrich *et al.*, 2010; Pan *et al.*, 2006; Cik *et al.*, 2014). Moreover, mobile phone data is suitable for pattern analysis due to large sample sizes. Mobility patterns (Calabrese *et al.*, 2013; González *et al.*, 2008) were analyzed as well as patterns in urban road usage (Wang *et al.*, 2012). Finally, activity location identification was performed based on mobile phone data (Chen *et al.*, 2014).

Several studies have comparatively investigated data quality. For instance, studies have compared GSM data with GPS trajectories (Iovan *et al.*, 2013; Hoteit *et al.*, 2014; Smoreda *et al.*, 2013). In addition, sociological aspects of mobile phone usage have been investigated, for instance with regard to analysis of places relevant in transport science (Licoppe *et al.*, 2008). Mobile phone billing data was utilized several times to obtain an OD-matrix, which can be done without much effort, because mobile phone billing data consists of space-time points. An early study in England (White and Wells, 2002) showed that the accuracy of billing data is not good enough to compute a reliable OD-matrix. Therefore, other researchers combined this data sources with others to get better results for OD-matrices. Some of these additional sources are signalling data in a Spanish region (Caceres *et al.*, 2007) and the Ile-de-France (Bonnel *et al.*, 2015), smartphone application data in Sweden (Mellegard, 2011), geo-spatial data together with census data in cities in USA, Portugal, Brazil (Toole *et al.*, 2015). Most of the findings suggest that mobile phone billing data can be a good proxy for overall tendencies of human mobility, thanks among other things to the large samples of persons and days involved. Altogether, GSM data is a powerful tool for analyzing human mobility (Song *et al.*, 2010) as it is shown by the increasing interest of researchers. Nevertheless, usually further data sources are needed to get reliable results. This chapter shows why estimates based on mobile phone billing data are valuable in case of long-distance travel demand.

Mobile phone data and national travel surveys have only been sporadically compared so far. Bekhor *et al.* (2013) conducted an evaluation in Israel. However, the study sample was comparatively small in terms of person-days.

The focus was not on longitudinal travel behavior, and the first data source preceded the second one by 10 years (with a 25% population increase). Similar work has been done in the USA (Huntsinger and Donnelly, 2014), but was also limited to a regional level (North Carolina). Neither of the two studies provides statements about long-distance travel demand, since they focus on other aspects of travel behavior. This gap is closed in the remainder of this chapter.

6.2 Data Sources

6.2.1 Mobile Phone Billing Data

The study described in this chapter is based on an anonymised mobile phone billing data set recorded by Orange™ France. It consists of Call Detail Records (CDRs) covering the mobile phone usage of around 23 million users of the Orange™ network in France during a period of 154 consecutive days (13 May 2007 to 14 October 2007). Given a population estimate of 63.9 million inhabitants in 2007, that is roughly 35.9% of the French population. The population estimate is the average of the monthly estimates for the period between May and October 2007 obtained from the French National Institute of Statistics and Economic Studies website (`www.insee.fr`). The numbers correspond with estimates made by Orange™ that mobile phone penetration in France in 2007 was 86% (ARCEP, 2016), and with the estimated market share of Orange™ in that year (43.5%).

Each CDR contains information about an action (outgoing/terminating call or SMS) which took place in the network. The information needed for our purpose is the caller ID, the time and duration of the action, and the Base Transceiver Station (BTS) that was the connection point for the mobile phone at the start of the action. A BTS is responsible for the wireless communication between the network and the mobile device. Several BTS can be located on a single tower serving different directions and/or technologies. The location of every tower is known. Given information on the location and time of each action, individual users can be traced and their movements can be extracted. The accuracy of reconstructed movements depends on the frequency of actions since no information on the phone in idle mode is given in the data.

The CDR data set has several limitations. First, the action frequency is

comparably low, because mobile data usage was not as intense in 2007 as it is today. Second, the data set does not cover a full year. Thus, any estimates for the missing time periods must be supported with complementary data sets. In addition to temporal inaccuracy due to the low call frequency, there is also spatial inaccuracy. The spatial information gained from CDR data is limited by the positions of the mobile network towers handling the BTS. For less densely populated areas of the country, a BTS can be several kilometers away from the actual position of a mobile phone. Finally, no information about phone calls made abroad is available in this data set. Even though it is known that France has one of the highest ratios of domestic trips to trips abroad within Europe (OECD, 2012; Eurostat, 2016) this circumstance limits the range for which valid estimates can be made. It is accounted for this limitation with respect to the special situation of a large central European nation in the results section below.

It has been shown that mobile phone billing data should be used with caution when analyzing mobility (Ranjan *et al.*, 2012). Nevertheless, most limitations do not have a substantial impact when focusing on long-distance travel demand. The spatial and temporal inaccuracies described above are relatively small since the scope of this work is on large spatial and temporal scales. Additional signalling data as it is used in Bonnel *et al.* (2015) would improve the quality of the results since it offers more frequent data, but is not available in this case. Signalling data is an additional information recorded by the network companies, e.g. when an idle mobile phone leaves specific pre-defined areas. Still, mobile phone billing data can provide a lower bound to the actual value. When comparing CDR data with survey data, it has to be accounted account for the missing roaming data and focus on the national travel. A detailed discussion on the limitations of the data and the methodology can be found in Section 6.5.

6.2.2 Survey Data

The results of the CDR data analysis were compared to a national travel survey. The *Enquête Nationale Transports et Déplacements (ENTD)*, the French national travel survey, is taken for this purpose. The ENTD is conducted every 10 to 15 years (1967, 1974, 1982, 1994, 2007–08). Various actors are involved in the ENTD, including the French Ministry of Transport, INSEE (the French National Institute of Statistics and Economic Studies)

and IFSTTAR (French Institute of Science and Technology for Transport, Development and Networks). The most recent one was conducted from April 2007 to April 2008, and most parts are publicly available (IFSTTAR, 2016). The ENTD 2008 was used here, because it includes the time period covered by the CDR data described above. Moreover, one of its goals was to analyze long-distance mobility which is advantageous because comparisons between the two data sources in terms of long-distance travel behavior were made possible.

However, the sample size of the ENTD 2008 is much smaller than the available CDR data. In total, 20,178 households and 44,958 individuals were surveyed. Just 18,632 (representative) individuals were chosen for the long-distance travel module of the survey (Armoogum *et al.*, 2008). The latter were asked to report their long-distance travel practices within the preceding four weeks as well as the preceding 13 weeks. Detailed information on tours taken during the four-week period is publicly available, while just the absolute number of tours is available for the 13-week period. Consequently, results of the 13-week questionnaire are only used for a comparison of the total number of trips, while the four-week data was used for further analysis like distance distribution. In the ENTD 2008 a long-distance journey is defined as either a journey with the furthest destination more than 80 km from home as the crow flies or 100 km actual travel distance. Journeys, which include at least one overnight stay, were added in the survey, but were removed from the data set for this work in order to compare the actual long-distance travel behavior. It is accounted for the differences between the data sources in the comparisons (see Section 6.4). Unless otherwise stated, the four-week ENTD records were compared to the CDR data due to the lack of detailed 13-week records.

6.3 Methodology

The mobile phone billing data set described above was far too big to be analyzed completely within the framework of this thesis. The reason is the limited server access time that was granted for this work in combination to the computational heavy tour extraction algorithm that is described in detail in Subsection 6.3.3. Therefore, the number of mobile phone users and their CDRs studied is limited. Two selection steps were performed.

First, a set of municipalities was chosen. Second, from each municipality a subset of customers was selected in order to investigate their travel behavior. Even though the analyzed data set is just a sample of the whole data set, the sample size analyzed here exceeds by far the size of any data set collected with a traditional survey.

The impact of population size on the long-distance travel of the cities' residents has to be analyzed since the German literature suggests that inhabitants trade off daily travel against more long-distance travel (Holz-Rau *et al.*, 2014; Schlich, 2001). In the following sections, both selection processes are described in detail, and the algorithm used to extract the long-distance tours from the mobile phone data is presented.

The following definitions will be used henceforth in this chapter:

- Home environment: The area within a radius of 80 km from the home location.
- (Home-based) tour: A chain of activities and trips starting and ending at the home location (sometimes referred to as a 'journey').
- LD tour: A tour which leaves the home environment and therefore is a long-distance tour, because the destination is at least 80 km away.
- LDF tour: A domestic LD tour. Thus, an LD tour with a destination within France.

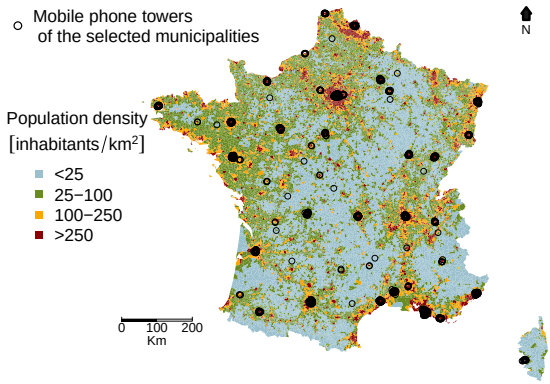
The focus of this chapter is the analysis of LDF tours due to the limitation of the CDR data of missing roaming information.

6.3.1 Municipality Selection

As described above, the number of tracked persons in the CDR data had to be limited. As a first step, a set of municipalities was chosen and further analysis was focused on the inhabitants of those municipalities. The municipalities were selected such that they are well distributed, spatially as well as in terms of size of population. In the end, every major city was selected and a random sample of smaller communities was added to the selection.

All mobile phone towers and their Base Transceiver Stations (BTS) within the chosen municipalities were identified. Each mobile phone tower can hold several BTS (serving different directions and/or technologies). Furthermore, several towers can be at the same location, e.g. on top of the same building. The final selection of mobile phone towers is shown

Figure 6.1: Mobile phone tower locations of the selected municipalities



in Figure 6.1. In total, 23,438 Base Transceiver Stations in 3,631 distinct locations served the chosen municipalities. There can be several BTS at the same location for two different reasons. Either there is one tower operating several BTS, or there are several towers at the same location (e.g. one for each technology). These BTS cover the 58 municipalities chosen for analysis. They were used to identify the inhabitants of the municipalities. Bold circles indicate where there are many towers in close proximity. This was the case in dense cities. The cities located closest to a border are Calais (on the coast), Lille, Strasbourg and Mulhouse. It is expected that the limitation to domestic travel will reduce the number of observed long-distance tours substantially in these cities. Furthermore, all regional centers (identified by high population densities) were included in our selection.

6.3.2 Selection of Tracked Persons

In order to decide whether a customer was an inhabitant of one of the municipalities considered one needs to infer the customer's place of residence. An analysis of home anchors (Ahas *et al.*, 2008b, 2010) was undertaken for this purpose. Anchors are the mobile network towers which were most frequently used by a customer during a specific time of day. To compute home anchors the focus was put on nighttime hours (9 p.m. to 6 a.m.), because most people are expected to be home for the majority of nights. An additional requirement was needed to avoid wrongly setting a home anchor by the call actions of a single night. Thus, it is determined that a tower is a home anchor candidate only if the phone was in use at that location for at least seven distinct days in a month. Following these rules, home anchors were computed for each customer and for each of month. Hence, each customer had up to six home anchors. Many persons did not have an anchor for May and October since these months were just partly covered by the CDR data. A monthly analysis was performed in order to identify persons that relocated their home (e.g. to a summer house).

For around 18 million users there was at least one month when it was possible to identify a home anchor. A customer was considered to be a resident of a municipality if he or she had at least three home anchors within the given municipality. This threshold was chosen because there were just half a month of observations during two of the six monitored months. Thus, there was a substantial share of customers who did not have home anchors in those months. Therefore, most of the persons had just four home anchors. Hence, it is assumed that people lived at a place if they had three quarters of their potential home anchors at the same location. All customers who were inhabitants of any of the selected municipalities were chosen for further analysis. This subset contains more than 1.4 million customers and, therefore, captures over 17% of the population of the selected municipalities.

An algorithm was applied to identify machine-to-machine devices. Such machines are SIM-card devices that are not used by humans but are automated. One can detect these machines by looking for specific periodic behavior. This behavior is relatively easy to identify since machine communication follows pre-defined rules. For example, a device communicating with a specific other device over a series of days at the same time of day is likely to be a machine-to-machine device. All of the identified machines

Table 6.2: Number of tracked persons by size of municipality

Population [in 1,000]	Tracked persons	Number of municipalities
Paris	4,953	1
200–900	19,394	10
100–200	25,294	13
50–100	9,580	5
20–50	7,461	4
10–20	7,730	5
5–10	3,190	5
1–5	1,376	7
rural (<1)	896	8
Total	79,874	58

were removed from the subset. As a consequence, the size of the subset of customers was reduced to 1.39 million.

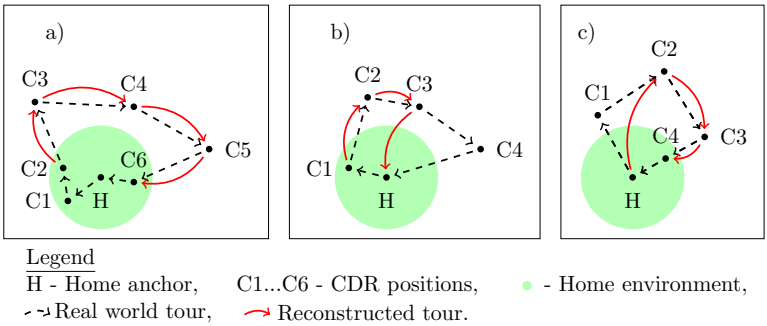
In order to identify the persons who actually made at least one LDF tour, an additional filter had to be implemented. A single month (June 2007) is chosen and investigated whether the persons left their home environment during that month. More than 814,000 of the identified residents did so. Of those, a subset of persons is randomly selected for a detailed analysis of their long-distance travel behavior. In total, 5,000 residents of Paris, 2,000 individuals from the other major cities and all identified persons from the smaller municipalities were chosen. After additional data cleaning 79,874 persons were left, and their long-distance travel behavior was studied. Table 6.2 shows the number of persons and municipalities by population size.

6.3.3 Identifying Long-Distance Tours in CDR Data

Unlike surveys, mobile phone data does not directly provide information about tours undertaken. The available information is a series of time-space points. It has been shown how the series can be used to infer the home locations of mobile phone users. In the following, the extraction of long-distance tours is described in detail.

When scanning the users' CDRs it is supposed that an LDF tour started

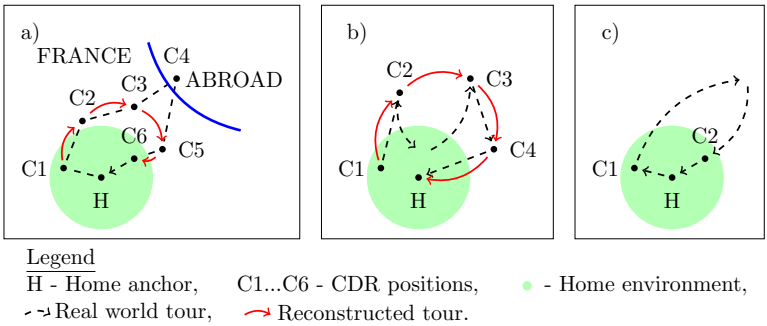
Figure 6.2: Visualization of the tour reconstruction algorithm a) Perfect tour reconstruction, b) Tour with unobserved end: C4 is after 14 Oct, c) Tour with unobserved start: C1 is before 13 May.



every time a CDR with a location outside the home environment occurred following a CDR located within the home environment. The tour was assumed to end with the first CDR back in the home environment. A sketch of a single construction process can be found in Figure 6.2a). The initial situation consists of the home anchor (H) and the home environment (green circle). The locations of the CDRs are then identified as C1, C2,...C6, whereby their sequence is given by their numbers. The black dashed arrows show a possible path of the user, while the red solid arrows form the reconstructed tour. In the sketch in Figure 6.2a) the reconstructed tour fits the initial real-world tour quite well. This is not always the case. A problem is the boundary of the traced time period. Tours that had not finished before the end of the observed time period had to be truncated without any information on their further duration (Figure 6.2b)). Likewise, tours that started before the recorded time had to be truncated (Figure 6.2c)).

Moreover, the character of CDR data caused further limitations. First, there is no information about mobile phone usage outside of France. This lack of information led to wrongly inferred final destinations during tour reconstruction (Figure 6.3a)). Without any mobile phone activity between the home environment and the border, even an around-the-world tour would

Figure 6.3: Failed tour reconstructions: a) Missed tour abroad, b) Two tours merged c) Tour completely missed



be missed. This was likely the case for most of the international tours. Second, low-frequency mobile phone users could go on two distinct tours without any mobile phone activity within the home environment between them. In such cases, the tour reconstruction algorithm merged the two tours due to the lack of a separating CDR (Figure 6.3b)). Third, the worst case is a user without any CDRs related to his or her long-distance travel. Without CDRs indicating an exit of the home environment, no tour could be reconstructed (Figure 6.3c)). This is the most critical and likely the most frequent reason for a failed tour reconstruction. In addition, it was also possible to miss certain parts of a tour or its final destination. Note that all limitations led to a lower number of tours in comparison to the real world. Therefore, it can be assumed that the number of LDF tours identified by the algorithm is a lower bound of the total.

It is possible that one person had two devices, e.g. a business phone and a private phone. Therefore, a duplicate check was performed. It was checked whether there were two customers with home locations close to each other (less than 500 m) and had similar travel behavior. Similar travel behavior occurs, if more than 75% of the LDF tours overlap in time and had a close destination (at most 5 km deviation). Just 34 duplicates were found. One of the duplicates was removed from each pair of look-alikes.

6.3.4 Accuracy of the Estimates

The number of tours (per capita) is the main indicator analyzed that describes long-distance travel demand. Therefore, the accuracy of this measure is discussed here. There are several reasons to assume that the number of long-distance tours reconstructed from CDR data is still lower than the actual number of tours. First, it is very likely that many tours were merged (see Figure 6.2b)) or just not identified (see Figure 6.2c)). Especially the latter is assumed to lower the actual number of tours substantially since persons do not always place a call, when they leave the home environment. Second, it is assumed in the remainder of the chapter that persons, which did not do a long-distance tour in the reference period (June 2007), did not perform any long-distance tour. This is due to the customer selection. Again, this assumption leads potentially to a substantial under-reporting of the number of long-distance tours. In contrast, the focus on domestic tours might lead to a slight over-estimation of domestic long-distance tours, because an international tour might be recorded in the CDR data and, therefore, counted as domestic tour. However, this is not assumed to happen very often. In addition, the effect of the under-reporting is likely to be much higher than the latter opposing effect.

6.3.5 Seasonal Tour Frequencies in Survey Data

Analysis of CDR data was performed in comparison with the French national travel survey. Therefore, the survey data had to be adjusted in order to make the two data sets comparable (.g. international journeys had to be excluded). A major difference between the available CDR data and the survey data is the time periods covered. While the French national survey covers a whole year, the CDR data is limited to five months (mid-May to mid-October). Consequently, the share of tours within these five months had to be computed for the survey.

A detailed analysis of the tour frequency distribution in the ENTD was performed. For each day of the year, the number of tours and the number of persons reporting for the given day were computed. Subsequently, the number of tours was summed up and scaled by the number of respondents. Our computation shows that around 46.2% of all LDF tours took place within the five summer months. The share is higher than 5/12. Thus, the assumption that people tend to travel more during summer is confirmed. It

is accounted for the higher share of journeys in the summer in the following section.

6.3.6 Indicators for Analysis of Long-Distance Travel

Several indicators of long-distance travel demand are analyzed in the following section. While these indicators usually are reported directly in a survey, they have to be calculated in a CDR-based data set. It is shown in the following which indicators were chosen and how they were extracted from the data:

- Tour distance: The distance of a tour is defined as the crow-fly distance between the home location and furthest point visited on the respective tour. Particular interest is the distribution of tour distances.
- Destination: Again, the furthest point is defined as the destination of a tour. Since this definition is not accurate analysis of destinations is limited to the regional level.
- Share of long-distance travellers: A sample of 30 days and 1.4 million persons was selected to analyze how many persons are long-distance mobile. In other words, the share of persons that did at least one long-distance tour in the sampled period is calculated.
- Tour rates of mobile persons: For those, who did long-distance travel, the number of tours per person is calculated.
- Long-distance travel demand: The share of mobile persons is combined with the number of tours of mobile persons to get an estimate of total number of long-distance tours per capita.

6.4 Analysis of the Data Sets

Information regarding long-distance travel behavior (e.g. tour distances, tour frequency) is obtained from mobile phone data and compared with the ENTD 2008. The main differences between the two data sets are pointed out in the following section. The most important result is that long-distance travel demand for a whole year is heavily under-estimated if one relies solely on numbers given in the ENTD 2008. All of the results presented in this section are limited to journeys within France with a destination of more than 80 km away from the home location.

6.4.1 Comparability of the Data Sets

Before the two data sets can be compared, it is necessary to discuss whether the data sets cover the same scope and, therefore, whether they should be compared.

The respondents of the ENTD 2008 survey cover a variety of socio-demographics. The survey analysts claim to have a representative sample of the French population after weighting. Nevertheless, the results of long-distance travel demand are based on the responds of 5,000 persons, namely those that actually claim to do long-distance tours. In contrast, the CDR data does not report socio-demographics. A random sample of 80,000 customers was drawn from a pool that covers 36% of the actual population. It was accounted for spatial distribution and population size. Due to privacy regulations it is not possible to get socio-demographics of Orange™ customers individually or for the whole sample. However, there is no reason to assume that mobile phone users (86% of the whole populations) or Orange™ customers (43% of mobile phone users) differ substantially from the whole population. In addition, it is known that respondents of long-distance travel surveys under-report their travel behavior. Thus, it is not clear that 5,000 persons with representative socio-demographics give results which are more accurate than 80,000 randomly drawn persons.

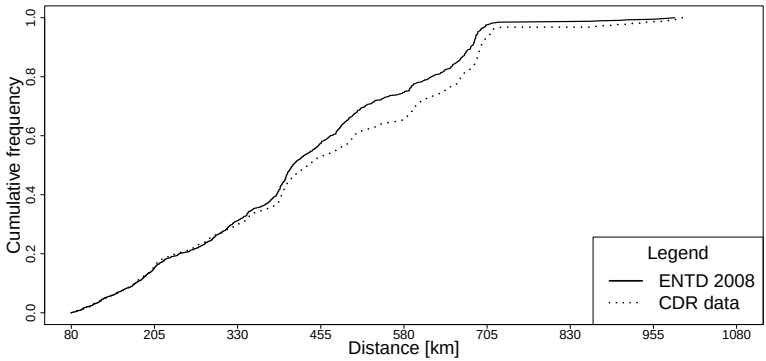
Therefore, a comparison is reasonable and valuable. Differences in the two data sets should not be treated as a fact. Likely, neither of the two data sets tells the ground truth. The following section with results rather indicates problems with survey-based long-distance travel demand data, shows that results should be treated with caution and illustrates that there is a need for alternative data sources.

6.4.2 Tour Distance Distribution

Distribution of LDF tour distances is of interest and is investigated here. For CDR data, the distance of a tour is defined as the distance as the crow flies between the home anchor and the furthest known point (mobile phone tower) away from home during this tour. In case of the ENTD, the respondents were asked to report the crow-fly distance to the main destination.

This subsection focuses on the inhabitants of a single municipality since the LDF tour distance is dependent on the location of the home (e.g.

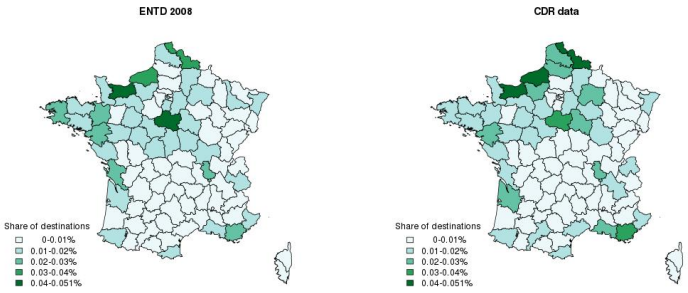
Figure 6.4: Cumulative frequency of LDF tour distances for Paris residents



based on distance to the border, distance to the next big city, surrounding sea/mountains, etc.). Paris was chosen for this analysis, because it is a city that is well represented in both data sets.

The results are shown in Figure 6.4. Cumulative frequency of tour distances is compared between both data sources. One can see that the CDR data almost perfectly reflects the survey data for residents of Paris. In the range of 450–650 km, the ENTD reports a slightly higher share of LDF tours than the CDR data. This may be explained by the respondents' underestimation of travelled kilometers for very long tours, i.e. tours of around 1,000km (see Wolf *et al.* (2003)). Using the Kolmogorov-Smirnov test to compare the two distributions shows that the distributions are significantly different (p-value $< 10^{-15}$). This is not surprising since these kind of tests are very sensitive for variations around the mean, which is the case here. In addition to the Kolmogorov-Smirnov test, the scope was sub-divided in 25km-bins and a Chi-Square test was performed. The Chi-Square test shows that the hypothesis that the two discretized distributions are drawn from the same main distribution can not be rejected (p-value = 0.29). This result indicates that the two data sets cover the same travel patterns.

Figure 6.5: Destinations of Ile-de-France residents (departement level)



6.4.3 Trip Distribution

Trip distribution is an important part of travel demand models. Therefore, trip distributions in the two data sets are compared. It is assumed that the furthest point on a tour in the CDR-based data was the main destination of a journey starting at the home location. The survey reports the main destination of the home-based tour. The difference of the two definitions is not expected to have large influence on a big scale. This analysis was limited to the residents of the Ile-de-France region, which covers the metropolitan area of Paris. This limitation was necessary since this is the only region, which is well represented in both data sets. Additionally, the survey was limited to the months May to October in order to avoid seasonal effects in the analysis.

Figure 6.5 shows the distribution of the destinations in the two data sets. The analysis was performed on the departement level which is second highest administrative level in France. The destination distribution is similar in the two data sets. Areas that are frequently visited are the close Atlantic coast, the Cote d'Azur, Lyon and the area to the south of Paris. Less frequently visited are the Bretagne, the southern Atlantic coast and the surroundings of Paris. The rest of France does not play a big role as a destination of domestic tours for residents of Ile-de-France. However, also

Table 6.3: Number of persons performing LDF tours

	CDR Data	ENTD	ENTD
Reporting/Tracked Interval	30 days (June)	28 days	91 days
Surveyed persons	1,388,941	18,632	18,632
LDF mobile persons	814,381	4,796	8,743
Share of LDF mobiles	58.6%	25.7%	46.9%
Selected for further analysis	79,874	4,796	8,743

small differences between the two data sets can be observed. Frequently visited departements seem to have even a higher share in the CDR data than in the ENTD. This variation might appear due to a much larger sample size of the CDR data. Nevertheless, the trip distributions are comparable, which is also confirmed by a statistical analysis of the destination patterns. The shares of visited destinations were transformed into a vector, where the i-th entry of the vector equals the share of visitors in departement i. The cosine-similarity of the two vectors for the two data sets has a value of 0.94 confirming that the destination patterns of the two data sets are very similar.

6.4.4 Share of Long-Distance Travellers

The number of long-distance travellers is a major question in travel demand modelling. Survey respondents are known to under-report their long-distance tours due to the high response burden of the corresponding items. Therefore, the CDR data was investigated with respect to the share of long-distance travellers. Because of the enormous amount of data, this analysis was restricted to a single month (June 2007). The results and the corresponding values in the ENTD 2008 are shown in Table 6.3.

Table 6.3 shows that the share of long-distance travellers within one month of CDR data (58.6%) is more than twice the share of travellers reported in four weeks of the ENTD 2008 (25.7%). While the 13-week reports of the travel survey show a higher share of travellers (46.9%), the value is still lower than 59%, as given in the CDR data, and lower than 61%, as estimated by Weckström-Eno (1999). The results support the assumption that a substantial number of survey respondents did not report

their long-distance journeys. Consequently, long-distance survey practice should not only pay attention to response rates, but should also find a way to convince respondents to report their journeys.

6.4.5 Tour Rates for Mobile Persons

It has been shown that the number of persons reporting LDF tours is much lower in the ENTD survey. The next question is whether the tour rates for those who reported tours also differ between the two data sets. The number of tours that took place within three months is compared (in case of the ENTD, the reported interval is 13 weeks). Figure 6.6 shows histograms for the two data sources. One can see that most of the ENTD respondents reported just one tour in this period, and just a very small share of persons travelled more than three times. The CDR histogram suggests that many people made two, three or four LDF tours, and a substantial number of tracked persons travelled more than five times within three months. The Kolmogorov-Smirnov test as well as the Chi-Square test were performed in order to test the similarity of the two distributions in Figure 6.6. Both tests suggest that the distributions of the number of tours differ significantly when comparing the two data sets (both p-values $< 10^{-12}$).

Monthly tour rates are compared as well in order to identify seasonal effects that might have had an influence. The monthly LDF tour rates for mobile persons are shown in Figure 6.7. The tour rates are substantially higher in the CDR data. This confirms our assumption of under-reported tour frequencies in surveys. Two aspects must be mentioned: first, the reference intervals differed slightly. While the CDR data was cut into monthly chunks, the ENTD survey responses referred to a four-week period. Second, May and October were not fully covered in the CDR data set. Thus, the shown tour rates are likely lower than the actual ones.

6.4.6 Long-Distance Travel Demand

Lastly, the total long-distance travel demand was analyzed. The number of LDF tours per capita was calculated based on the CDR data as well as on the ENTD data, as shown in Table 6.4. The reported frequencies refer to tour rates in the period from 13 May to 14 October 2007, henceforth called the *summer period*. Three different data sources from the ENTD were used:

Figure 6.6: Histogram of the number of LDF tours for mobile persons

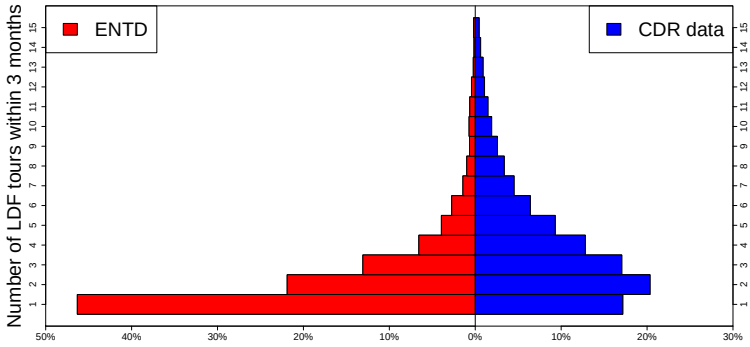


Figure 6.7: Average LDF tour rates per mobile person per month

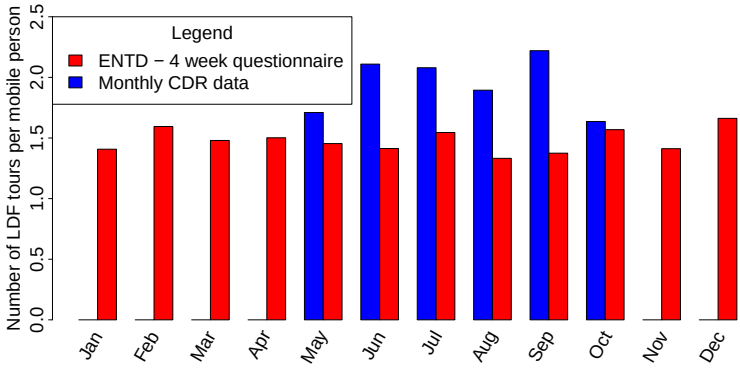


Table 6.4: Average number of LDF tours per capita (13 May - 14 October)

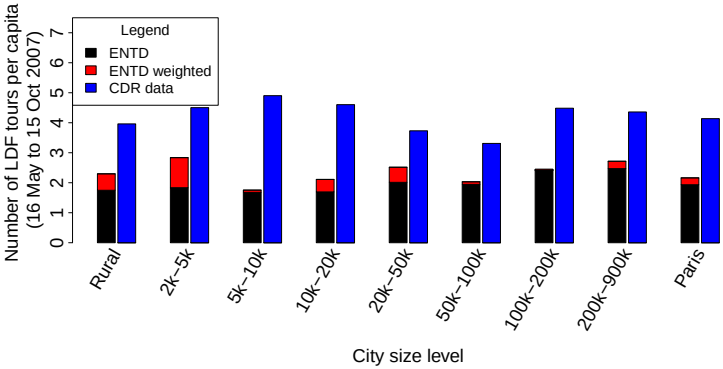
Reference interval	CDR data 5 months	ENTD 4 weeks	ENTD 13 weeks	weighted ENTD 1 year
Total	4.27	2.25	1.96	2.36
Compared to CDR data	100.0%	52.7%	45.9%	55.3%

first, the number of reported tours within four weeks; second, the number of reported tours within 13 weeks; and third, the number of projected yearly tours. For the latter the information that 46.2% of the yearly LDF tours were undertaken during the summer period and a weight provided in the ENTD is used to estimate the yearly travel demand.

One can see that the frequencies suggested by the ENTD are approximately half as high as those observed in the CDR data. Furthermore, adding the weighting factor proposed by the ENTD analysts does not change the main finding here. The factor of underestimation is much higher than is usually assumed (e.g. up to 1.3 in Cambridge Systematics Inc. (2013)). The seasonal effect has been taken into account, and there was also a spatial effect. Therefore, the LDF tour rates were analyzed according to the size of the home city in order to capture this effect (Figure 6.8). Additionally, the 95% confidence intervals are presented in Figure 6.9. Due to the smaller sample size, the confidence intervals of the ENTD survey results are wider than the intervals based on the CDR analysis. Once again, one can see that the CDR data suggests a long-distance rate that is twice as high as the ENTD survey outcome.

The differences described in this section have enormous impact. This can be seen in the resulting absolute numbers. Our analysis of the ENTD 2008 led to the assumption that the French population undertook 325 million long-distance tours per year. Limiting these to national tours and the summer period led to an estimate of 130 million tours. In contrast, the CDR data suggests that there were almost 240 million tours for the same population and time frame. It can be assumed that extending to the whole population, the full year and including international tours would not change the survey's under-reporting rate. Consequently, the ENTD

Figure 6.8: Average LDF tour rates per capita by municipality size for the summer period

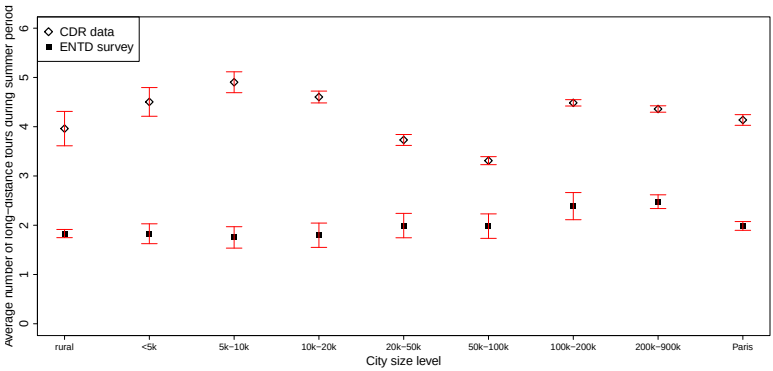


2008 under-estimates the annual long-distance travel demand by more than 260 million tours, with all of the attendant economic and environmental impacts. This number is based on simple scaling, assuming that the share of tours in the summer period and the share of national tours reported by the ENTD 2008 apply. Nevertheless, it gives an idea of the magnitude of the error. It has to be stressed as well that the tour frequency suggested by the CDR data is just a lower bound to the true value, which might even be higher.

6.5 Limitations

The methods used and the results obtained are discussed in this section focusing on limitations and their implications. Three types of limitations can be identified. First, the selection process of municipalities and persons can lead to a biased sample. Second, the reconstruction algorithm might be inaccurate, e.g. regarding the home location. Last, the data type itself, CDRs, has limitations, e.g. spatial inaccuracy. All types are discussed in the following.

Figure 6.9: Confidence Intervals of LDF tours per capita



A bias is likely to occur during the person selection. Frequent callers are more likely to be selected, because the home anchor algorithm needs frequent calls to identify a home location. Frequent callers more likely have more long-distance trips due to higher education, higher income, etc. Nevertheless, the sample was drawn from a pool of customers that captures more than 17% of the actual population. Considering the market share of 43% and the random selection, the sample set is assumed to have small bias, if any.

Furthermore, it is not clear whether the home location computation is precise. The computed home location can have a small spatial error (e.g. the neighboring tower of the actual home was chosen). This error can be neglected since the effect on the computation of (the number of) long-distance tours is rather small. The computed home location might also be far away, e.g. a shift worker is assumed to have his home at his actual work place. In this case it is likely that the number of reconstructed long-distance tours is lower than the actual number, because all of his non-work activities are clustered to a single tour.

Moreover, the question arises whether the municipality selection is representative. The set of municipalities captures all regions of France. In addition, all levels of population size are covered. All major cities are part

of the sample. Thus, the concerns are limited to the smaller municipalities. These were chosen randomly and are spread in space. The influence of the distance to the closest border was investigated. It could not be shown that it has significant impact on the number of domestic tours. There is probably an impact on the number of international tours, but this could not be checked here.

Cleaning of the data does not induce a bias by removing a large subset of usable data. The total share of removed customers is less than 1%. This includes the identification and removal of machine-operated devices, duplicated devices (one person with a private and business device) and network towers with corrupted location information.

CDR data has a limitation which usually makes it difficult to use for travel behavior analysis, namely spatial inaccuracy. In rural areas, Base Transceiver Stations cover big areas plus mobile devices are not always connecting to the closest tower. These facts lead to a potential error of several kilometers in the estimated distance of a long-distance tour. In case of long-distance travel behavior analysis, an error of a few kilometers is minor and can be neglected. Another issue related to the spatial inaccuracy has to be considered as well. Tours with a distance slightly above 80 km might not be found, because the location of the tower is closer to home than 80 km. On the other hand, the opposite effect occurs with similar probability. Tours slightly shorter than 80km are not a long-distance tour by definition, but might be identified as such. It is likely that these opposing effects cancel each other out.

One obvious limitation of the analyzed CDR-based long-distance travel behavior is the lack of international tours. In case of Lille or Strasbourg, it is likely that a substantial number of long-distance tours are missing. The reference data, the ENTD, reports that 11.5% of all long-distance tours from the departement Bas-Rhin (the departement containing Strasbourg) were international tours. However, international tours were excluded in this chapter. It has been also accounted for the fact that the CDR data was limited to 5 months covering the summer. A monthly analysis of the data has been performed in order to identify potential seasonal effects.

The second major limitation of CDR data is the low frequency of CDRs. Consequently, there are long-distance tours which can not be identified in the data. Short tours (in terms of duration) are especially affected since the probability to produce at least one CDR outside of the home environment

is low for short tours. However, this issue does not lead to over-estimation. Rather, the resulting number of long-distance tours is in fact a lower bound to the actual number which is expected to be substantially higher.

Finally, it is assumed in the analysis presented that LDF tours were only undertaken by those persons, who made at least one LDF tour in June 2007. This assumption lowers the number of LDF tours per capita as presented in Table 6.4. Adding tours of other mobile persons will lead to a higher total number of long-distance tours per capita which is again an indication that the actual number of long-distance tours is substantially higher than the tour rates presented in this work.

6.6 Discussion

The previous section has shown various limitations of the data analysis described in this chapter. It is important to discuss these limitations and their impact on the results in order to value the key findings. The impact on the analysis of the number of long-distance tours per capita is discussed in the following since this is the most important result indicating that the survey is heavily underreporting this rate.

First, there are limitations that were taken into account during the analysis. Analysis was focused on domestic tours in both data sets, because no international tours can be identified in the CDR data due to missing roaming information. It was assumed that there are no international tours in the CDR-based data. The error resulting from this assumption is likely to be very small. This is supported by the fact that just 10% of all tours in the ENTD 2008 were international tours. Hence, the main result, namely the under-reporting of long-distance travel, does not change, even in the case of falsely counted international tours.

The potential seasonal effect was eliminated when comparing the travel rates. In addition, mobile devices with similar travel patterns were identified and removed in order to avoid customers with two mobile phones in the data set. Therefore, these constraints can be neglected.

Second, there is a concern whether the analyzed sample is representative. Both, municipality selection and customer selection can not be proven to be representative. In case of the customer selection, the high market share of Orange™ , the large sample size and the random draw ensures that the

sampling error can be assumed to be rather small. In case of municipality selection, the selection is representative in terms of spatial distribution and size. Furthermore, the large sample size limits the bias related to other attributes, if there is any.

In addition, there are further limitations either with small impact, e.g. the home location algorithm, or with opposing effects that cancel each other out, e.g. the spatial inaccuracy of CDR data.

Summing up the discussion above, most of the limitations were either taken into account or are assumed to have small impact. This does not hold for the two issues discussed in the following. The assumption that persons do not travel at all if they did not travel in June 2007 has been made for the CDR data. Moreover, the frequency of CDRs is low leading to just few observations per day per person. These two facts are very likely to be responsible for a substantial under-estimation of the LDF tour rate. It is also very likely that the impact of these two facts dominates all other effects presented so far. Therefore, it is a valid assumption that the under-reporting factor estimated in this chapter is just a lower bound but is actually even higher. This finding is relevant for transportation research since it is a justification for further development of alternative data collection methods for the analysis of long-distance travel demand.

The number of annual long-distance tours per capita for the Orange™ customers has to be compared with the values measured in previous literature (Table 6.1). Scaling up the value calculated in Table 6.4 leads to an estimate of more than 9.0 annual tours per capita. Most of the literature reported in Table 6.1 has a lower number of tours, even though international tours were included in most of the studies. It is difficult to compare these values, because the studies cover different years and study areas. However, the relatively high number of long-distance tours reported in this chapter supports the request for investment in alternative data collection methods as mentioned above.

6.7 Conclusion

The long-distance travel behavior of the French population in the summer period of 2007 was analyzed in this chapter. The data source used was CDR data covering five months of mobile phone usage within the

French Orange™ network. It was found that the number of long-distance tours reported by the national travel survey for the same period was underestimated. The actual long-distance tour frequency was almost twice as high. Considering that CDR data gives a lower bound and that short tours (e.g. single-day commutes) were probably substantially under-reported, the long-distance tour frequency is likely even higher than shown in Section 6.4. Two reasons for the underestimated tour frequencies were identified. First, the average tour rates of mobile persons differed significantly between the ENTD survey and the CDR data. Hence, survey respondents underestimated their number of long-distance tours. Second, the number of persons reporting any long-distance tour was much higher than suggested by the ENTD data. Therefore, soft refusals were substantially responsible for the underestimated long-distance tour numbers. Consequently, alternative data sources are indispensable for a reliable estimate of long-distance tour frequency. Possible sources are either mobile phone data as presented in this chapter or extended GPS studies. Either way, inclusion of the device carried by most of us almost all the time -the mobile phone- is deemed necessary for a better understanding of long-distance travel behavior. The underestimation of long-distance travel has consequences for transport policy (e.g. wrong estimates of CO_2 emissions) and especially for the tourism sector (e.g. some markets are greater than assumed thus far). Finally, it also affects long-distance travel demand simulations as C-TAP and, especially, the initialization and calibration of these simulations. The following chapter shows how to utilize alternative data sources to generate a realistic synthetic population which can serve as initialization of a long-distance travel demand simulator.

Chapter 7

Population Synthesis

The previous chapter has shown the potential of alternative data sources to get better estimates for long-distance travel demand. However, turning a detailed data source into a realistic virtual population is not trivial. The main limitation using GSM or GPS data is the lack of socio-demographic information. This chapter proposes a new framework to generate a population using a big data source and fusing it with a traditional household travel survey.

This chapter is structured as follows. First, existing population synthesis approaches are reviewed and their limitations are described in detail. Then, the framework used to synthesize a virtual population for C-TAP is described. This is followed by a description of the application of this framework to actual data. The findings presented in this chapter build on work presented in Janzen and Axhausen (2017c).

7.1 Synthetic Population Requirements

A synthetic population is a crucial part of input for any agent-based simulation. It has direct and indirect impact on the travel demand simulated. Therefore, the population has to be as detailed as possible and close to the real world it represents. A synthetic population has to include at least the following information to be a suitable input for C-TAP. First of all, a population is built of agents. Basic socio-demographics (e.g. age, gender) is needed as well as the area of residence. The core of agents in C-TAP is defined by targets. In order to define targets, an estimation of the long-distance travel demand is required. The travel demand, i.e. the number of desired long-distance tours, has to include travel purpose and

desired duration. Given agents with their respective targets, C-TAP can simulate the long-distance travel demand. A realistic population has to also meet marginal conditions. In other words, the overall sum of tours and the share of all socio-demographic information has to match given counts. Before explaining a framework generating a population with the requirements described, known population synthesis approaches and their limitations are discussed.

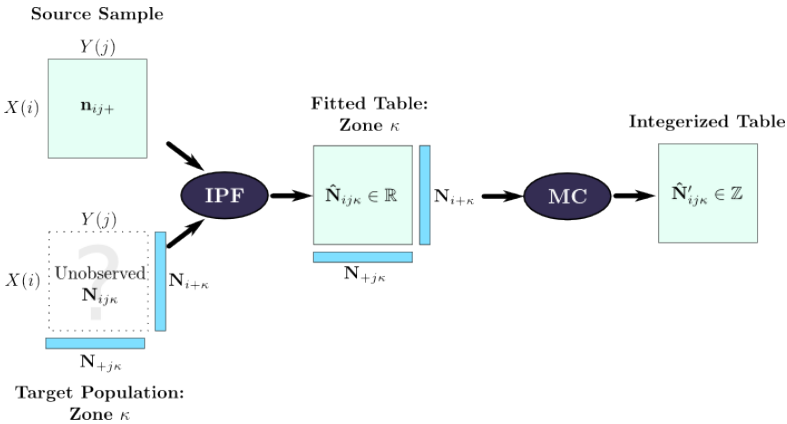
7.2 Limitations of known Approaches

Given a single data set (e.g. travel survey, smart-card data or GPS data) describing travel behavior, several approaches are known to generate activity schedules and travel patterns. The techniques used vary and include, but are not limited to rule-based approaches (Miller and Roorda, 2003), clustering (Ordóñez Medina, 2018) or various types of Bayesian Networks, like Hidden Markov Models (Anda and Ordóñez Medina, 2017). However, these approaches do not account directly for any totals. Therefore, advanced population synthesis algorithms try to fuse several data sources to synthesize a population that is more realistic.

Usually, two data sources are used to synthesize a population for agent-based simulations. First, travel surveys are performed to obtain detailed information on the persons and their travel behavior for a sample of persons. Second, official statistics (register data) provide information on the marginal totals. A fitting algorithm is then applied to create a population that matches the travel behavior reported in the survey data as well as the marginal totals of the register data. The most popular approach is the iterative proportional fitting (IPF) algorithm as sketched in Figure 7.1. A sample of a population is fused with marginal totals of the target population to get a fitted table. An additional rounding step generates an integer table that represents a realistic population.

In case of long-distance travel behavior, these two data sources are not sufficient, because travel surveys under-report long-distance travel heavily as it has been shown in the previous chapter. Therefore, an additional data source is needed. Passively collected mobile phone data based either on GPS or on GSM is suitable for this purpose. Mobile phone data is helpful since the obtained information on long-distance travel behavior is more

Figure 7.1: Sketch of a traditional population synthesis technique (IPF)



Source: Pritchard (2008)

reliable than results from survey data. On the other hand, the available samples of phone data lack socio-demographic information. Thus, it can not fully replace travel diary data. Potentially, phone data can be substituted by any reliable source on long-distance travel behavior. Mobile phone billing data as described in the previous chapter is utilized here.

7.3 Framework

In order to overcome the drawbacks described, is is proposed to utilize three data sources in our framework. First, all respondents from the survey are treated as possible agents of a synthetic population. Socio-demographic variables that are of interest (e.g. sex, age class, education), the zone of the residence and the number of long-distance tours per purpose are taken from the survey. The previous chapter has shown that the number of tours reported in surveys is too low. Assuming uniform under-reporting, passively collected phone data is used to adjust this under-reporting of long-distance tours in the travel survey data. The adjustment preserves the dependency

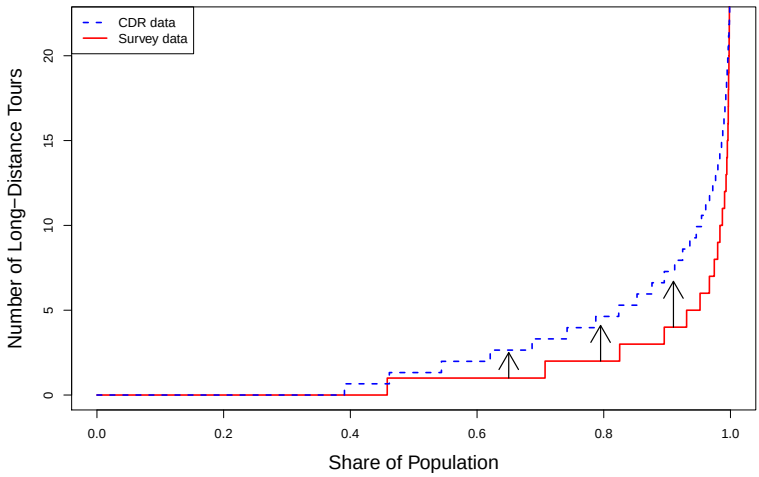
structure between socio-demographic attributes and the relative frequency of long-distance tours, using a technique akin to histogram matching in image processing. In the next step, generalized raking, a generalization of IPF, is used to find a weight for each person. Generalized raking can be used to compute weights for agents of a population and account for the marginal totals of the controlled variables. In this case, controlled variables are the number of tours per purpose (taken from phone data) and socio-demographics (taken from register data). The number of agents required for the population can be drawn with respect to the calculated weights. Finally, a duration needs to be assigned to each tour of the drawn agents. The duration can be imputed using a model based on the duration distributions given in the survey data. This model has to take into account zone types, purposes and some of the socio-demographic attributes. The result is a population of arbitrary size including information on socio-demographics and zone of residence as well as detailed long-distance travel description. The framework takes advantage of the knowledge of marginal totals from register data, reliable travel information from phone data and socio-demographic influences from survey data. Thus, it is more reliable than traditional approaches using only surveys and register data.

The framework presented above will be applied with three available data sources covering France in 2007. A national travel survey including a long-distance travel questionnaire was carried out at this time. In addition, register data is collected on a yearly basis by the National Institute of Statistics and Economic Studies (INSEE). Furthermore, a large sample of mobile phone billing data from Orange™ is available (see Chapter 6). These three data sources are combined following the framework described above in order to synthesize a population. In case of the phone data, the data set is scaled, because it covers just domestic travel within five months.

7.3.1 Histogram Matching

As mentioned above a major problem of long-distance surveys is the under-reporting of travel. However, survey data has to be the core of population synthesis for an agent-based simulation due to its detailed information like socio-demographics. Therefore, it is proposed to use any given long-distance travel survey and scale the number of long-distance tours while preserving socio-demographics and relative number of tours. The target

Figure 7.2: Illustration of Histogram Matching



distribution has to be provided by a reliable data source. A predestined data source for this purpose is big data, e.g. mobile phone billing data as presented in Chapter 6. A similar approach has been pursued by Fourie (2016) who synthesized a high dimensional data set from two-dimensional aggregate distributions.

The algorithm used for scaling is called Histogram Matching. Histogram Matching was initially used in the field of image processing, but is modified here. Technically speaking, the task is the following: Given two ordered distributions X_1 and X_2, find a monotonically increasing function f such that $f(X_1) = X_2$. The task is illustrated in Figure 7.2. The initial distribution is plotted as a solid red line and the target distribution as a dashed blue line. The two distributions shown here are based on the CDR and ENTD survey data presented in the previous chapter where the CDR data constitutes the target distribution.

The matching algorithm, a quasi Histogram Matching, comprises the

following steps: First, for each residential zone type, potential soft refusers are identified in the survey. Second, both distributions are ordered and the survey tour rate distribution is mapped to the target distribution given by the mobile phone data. Finally, the share of purposes is transferred. The fractions of purposes are kept and randomized rounding is performed in a later step.

The identification of soft refusers as well as the tie-breaking within the ordering of survey respondents needs additional information. For both purposes, a linear regression model is calculated which predicts the number of long-distance tours. Soft-refusers are identified as the respondents with the largest number of tours predicted by the model among those respondents who claimed not to travel for long distances. The tie-breaking procedure for the ordering follows a similar idea. The respondents are ordered by reported number of tours and by number of predicted tours in case of a tie. The scaling technique described here yields a population sample that includes socio-demographic information (coming from the survey data) and realistic long-distance travel demand (by matching targeted numbers from big data).

7.3.2 Generalized Raking

The histogram matching provides a set of persons with corrected travel demand values. However, a full synthetic population consists of a larger number of agents than provided by a survey. The well known Generalized Raking approach (Deville and Särndal, 1992; Deville *et al.*, 1993) is used to scale the number of agents. The idea of Generalized Raking is to calculate a population that is close to an existing weighted (sample of a) population and meets given marginal counts.

Technically speaking, the objective is to find a population $U = \{1, ..., k, ...N\}$, where each $k \in U$ has J attributes $x_k = (x_{k_1}, ...x_{k_J})$. Given are marginal totals $t_x = \sum_U x_k$, a sample of the population $S \subset U$ and an inclusion probability $\pi_k = Pr(k \in S)$. The inclusion probability is used to define starting weights $d_k = \frac{1}{\pi_k}$. Starting weights can be calculated using the Horvitz-Thompson Estimator, which is usually done in the studies of travel demand surveys.

The task is to find calibrated weights w_k for a distance function G, which

solve the system

$$\begin{aligned} &min \sum_S G(w_k/d_k) \\ s.t. &\sum_S w_k x_k = \sum_U x_k \end{aligned} \tag{7.1}$$

Assuming that G is positive and strictly convex, $G(1) = G'(1) = 0$ and $G''(1) = 1$ holds, leads to:

$$0 = g(\frac{w_k}{d_k}) - x_k \lambda \tag{7.2}$$

$$\Rightarrow w_k = d_k g^{-1}(x_k \lambda) \tag{7.3}$$

$$\Rightarrow \sum_U x_k = \sum_S d_k g^{-1}(x_k \lambda) x_k \tag{7.4}$$

where $\lambda \in \mathbb{R}^J$ is the Lagrangian multiplier and $g(v) = dG(v)/dv$. Equation 7.4 defines an equation system and a result of using Equation 7.3 and replacing w_k in Equation 7.1. After solving the equation system (7.4), one can easily calculate the calibrated weights w_k using Equation (7.3).

The complexity of the General Raking algorithm depends on the distance function G. A simple and prominent option is defined by the *linear method*:

$$G(v) = \frac{1}{2}(v-1)^2 \qquad v \in \mathbb{R} \tag{7.5}$$

which leads to $g^{-1}(u) = 1 + u$. Therefore system (7.4) transforms to

$$\sum_U x_k - \sum_S d_k x_k = \lambda \sum_S d_k x_k x_k^T \tag{7.6}$$

and the calibrated weights transform to $w_k = d_k(1 + x_k \lambda)$. Note that the so-called *multiplicative method* is equal to the Iterative Proportional Fitting (IPF) algorithm ($G(v) = v \log v - v + 1$, $g^{-1}(u) = exp(u)$).

7.3.3 Duration Model

The Generalized Raking as presented above yields a population including the number of activities that involve long-distance travel. Activity durations are not part of this activities. However, activity durations are required in

order to initialize the duration targets for C-TAP agents. The durations cannot be transferred directly from one of the data sources used above. A transfer would yield a non-continuous solution space, i.e. the resulting population would include just a small sample of diverse durations. There are two alternatives to tackle this issue. First, the solutions space, i.e. the set of valid activity durations, is cut in blocks. As a consequence the predicted variable is a duration block or a range of durations. This can be achieved using decision trees as it is done, for example, within *Albatross* (Arentze and Timmermans, 2004). Second, a continuous model can be used compute activity durations. Survival models like hazard-based models have not been employed often in activity duration prediction. Few studies include analysis of shopping durations (Bhat, 1996) or home-stay durations (Hamed and Mannering, 1993). However, this studies are limited to specific activity types and take advantage of additional knowledge. The main problem is that activity duration per se does not heavily rely on socio-demographics, especially in case of long-term activities as used in C-TAP.

Thus, a distribution-based imputation is implemented to add activity durations to the population. The algorithm for duration computation is composed of the following steps:

1. Take a long-distance travel survey and cluster the set of activities by activity type and few socio-demographics (like income and age).
2. For each of the clusters compute the (discrete) distribution of activity durations.
3. Find a continuous function for each cluster that fits the corresponding discrete duration distribution.
4. For each activity in the synthetic population do the following: find the corresponding cluster and draw a sample duration based on the given continuous distribution.

This procedure has two advantages. First, durations are sampled from a continuous distribution. Second, the duration distributions given by the survey are mirrored. On the other hand, just few socio-demographics can be used, because more variables used would yield more and smaller clusters. However, small clusters are prone to outlier effects and have to be avoided here. Nevertheless, socio-demographics are likely not to have large influence here as it was discussed above.

Applying the duration imputation method described in this subsection leads to a synthetic population that includes activities and their durations.

This can be easily transformed to a C-TAP population where activities translate to percentage-of-time, frequency and duration targets. An example of the duration imputation framework is presented in Section 7.4.

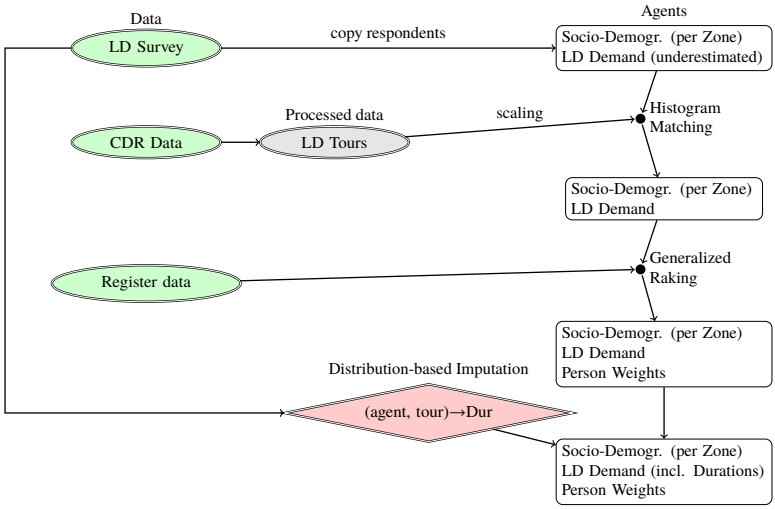
7.3.4 Illustration

The framework including all single steps is illustrated in Figure 7.3. The green ellipses on the left side of the figure represent the used data sources, namely a long-distance travel survey, CDR data (or a other equivalent big data source) and register data providing marginal totals. The set of agents and the detail of information available after each of the steps is shown on the right side of the graphic. The survey provides a set of agents corresponding to the respondents with socio-demographic information and the long-distance travel demand. However, the travel demand is underestimated due to soft-refusers. Therefore, CDR data is used to extract the actual long-distance travel demand. Histogram Matching as described in Subsection 7.3.1 is used to scale up the travel demand to a more realistic distribution. The resulting sample can be taken as a true sample of the target population. Consequently, it can be used as input for the Generalized Raking algorithm along with register data. Generalized Raking produces a synthetic population with realistic socio-demographics and number of trips/activities. In the last step, a duration is imputed to all activities of the synthetic population. The synthetic population resulting from this framework can be utilized in the C-TAP simulation to generate long-distance travel demand.

7.4 Application

As mentioned above, an application of the framework is given in the following. The data sources used in this example cover the same region (France) and the same time period (year 2007). The big data source is covered by CDR data originating in the Orange™ network and was introduced in Chapter 6 (see also Janzen *et al.*, 2018). The long-distance survey is the *Enquête Nationale Transports et Déplacements (ENTD)* of 2007/08 (IFSTTAR, 2016). The characteristics of the survey were given in Chapter 6 as well. Finally, total counts of socio-demographics are provided by the official website of *Institut national de la statistique et des études*

Figure 7.3: Population Synthesis Framework



économiques (INSEE) which is the national institute of statistics in France. The total count of the number of long-distance tours is taken from the CDR data since this is the most reliable data source for this purpose.

Table 7.1 shows a sample of nine agents (ID 1-9) from the synthetic population. The table also illustrates all steps of the synthesis. The left part displays the respondents of the survey and the information that was copied. This includes relevant socio-demographics, residence type, number of long-distance tours reported and the person weight d_k calculated by the survey analysts. The socio-demographics are limited to gender and age in this example. Further characteristics like income or car ownership are taken into account, but are not presented here for simplicity. The residence type is defined by the population size of the municipality the survey respondent lives in. The person weights are needed in a later step for the Generalized Raking approach. The next column provides the scaled number of long-distance tours which results after applying the Histogram Matching. As you can see most of the numbers grow due to under-reporting in the survey.

However, special interest should be paid to agents 8 and 9. Both agents are based on survey respondents that reported not to travel. Nevertheless, agent 9 is identified as soft-refuser using the model described in Section 7.3.1. Therefore, its number of long-distance tours is scaled up. On the other hand, agent 8 is likely to be to immobile in terms of long-distance travel. The agent remains to be part of the synthetic population in order to match the total counts despite the fact that it does not contribute to the travel demand simulated by C-TAP. The last column gives the new weights w_k as calculated by the General Raking. The distance function used for this calculation is $G = 0.5(v - 1)^2$ as described in section 7.3.2. The total counts used in this example are the following:

- Gender: female (32,050,119 agents), male (30,084,747 agents).
- Age bins: 0-15 years (11,369,872), 15-30 years (11,757,993), 30-40 years (8,341,114), 40-50 years (8,711,666), 50-60 years (8,327,709), 60 years and above (13,626,512).
- Population of residence [in $1,000$]: <1 (9,905,300), 1-5 (15,073,600), 5-10 (7,119,800), 10-20 (6,559,700), 20-50 (9,035,000), 50-100 (4,915,800), 100-200 (3,744,500), 200-900 (3,584,700), Paris (2,211,400).
- Number of long-distance tours: in total 464,538,704 tours.

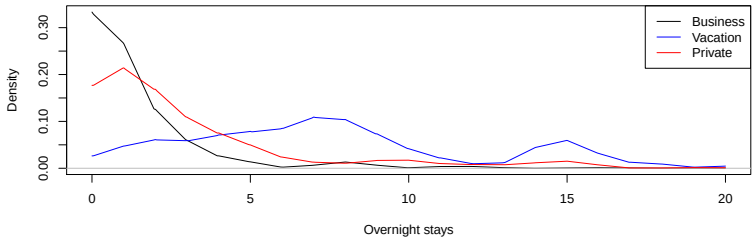
The total population is around 62 million agents which is the officially registered population size of France in 2007. Note that more than 40% don't perform any long-distance travel as it is shown in Table 6.3. These persons can be neglected in an agent-based simulation of long-distance travel demand.

The first agent in Table 7.1 performs eight long-distance tours per year and represents 4,463 virtual persons in the total 62 million-population with the same socio-demographics. Therefore, the next step in the population synthesis framework is to create 4,463 copies of this agent and sample a duration for each of the eight tours for each copy. The durations are sampled from a distribution that is taken from the survey and smoothed. You can find an illustration of the duration distribution by purpose for the first agent in Figure 7.4. The distributions meet general expectations. The business tours are usually short, namely just a few days. Private tours are a bit longer, but also mostly limited to one week. Vacation tours are substantially longer and have peaks at 7-day and 14-day durations which are the typical vacation durations. Note that these are durations specific for

Table 7.1: Sample of a population synthesis including single steps

	Step 1: Copy						Step 2: HM	Step 3: GR
ID	Gender	Age	···	Residence	#LD-tours	d_k	scaled #LD-tours	w_k
1	M	15-30	···	10k-20k	8	1221.1	8	4463.1
2	F	50-60	···	20k-50k	8	629.5	8	2086.8
3	M	40-50	···	Paris	3	2014.7	3	447.3
4	F	50-60	···	<1k	8	411.3	10	364.3
5	M	40-50	···	1-5k	12	1011.2	15	6211.1
6	M	50-60	···	50k-100k	4	1518.6	5	2136.2
7	F	40-50	···	100k-200k	8	2746.4	10	2870.9
8	F	40-50	···	10k-20k	0	2881.6	0	9577.3
9	F	>60	···	50k-100k	0	2144.1	1	3764.3
···	···	···	···	···	···	···	···	···

Figure 7.4: Duration density distribution for long-distance tours by purpose of person that is male and 15-30 years old



young male individuals. Other person types have different distributions, e.g. older persons tend to have longer business tours and less 14-day vacations.

7.5 Discussion

The population synthesis framework presented above is the first approach to generate a population covering long-distance travel demand. A traditional survey is combined with register data as well as mobile phone data. The synthetic population meets characteristics reported in the survey as well as socio-demographic totals and travel demand taken from big data.

Nevertheless, the population created using the approach presented has limitations. Household structure is not part of the synthesized population. In case of long distance travel, households are relatively important, because many long-distance trips are performed with all household members, e.g. vacation. A multi-level fitting algorithm, as it has been already used for synthetic populations (Müller, 2017), can be applied for this purpose. However, C-TAP is not simulating household decisions. Simulating joint decisions is proven to be a challenging task (Dubernet, 2017; Hackney and Axhausen, 2006) and is one of the main objectives for further development of C-TAP (see Section 10.2).

A drawback of the population synthesis framework is the amount of data that is needed. Three data sources are used to generate a population that describes individual long-distance travel demand. All data sets are collected by different institutions and need to be gathered in advance. First, register data is usually collected and provided by official statistics authorities. Second, survey data is collected by private companies or research institutes. Third, big data is collected by rather large companies, e.g. mobile network operators. The first two data sets are mostly publicly available or shared for research purposes. This does not apply to the latter data sets. Due to several reasons, like privacy issues or the companies' need for revenue, big data is rarely available. Nevertheless, all three sources are necessary as it has been pointed out above.

Furthermore, several characteristics of long-distance travel are not part of the population. This includes travel time values, destinations and modes. Some of these attributes are usually modelled in synthetic populations (e.g. Müller (2017)). However, destination choice and mode choice is part of the decision making within C-TAP. Therefore, this information is not needed as an attribute of the agents.

Chapter 8

Application

Theoretical background of C-TAP as well as data collection and processing has been presented in previous chapters. This chapter aims to illustrate that long-distance travel demand generated by C-TAP is valid and, therefore, can be used for policy analysis in the future. The remainder of this chapter is divided in two parts. First, the scenario that is used in this application is described in detail. Second, the simulation output is evaluated.

8.1 Scenario Configuration

The scenario used to demonstrate the concepts of continuous target-based activity planning in case of long-distance travel covers the Ile de France. Ile de France is a region in France that includes Paris and is home to more than 12 million persons. The long-distance travel demand of the whole population is simulated for a year in this application. Further details of the scenario definition are given in the remainder of this section.

8.1.1 Population

An important part of the scenario initialization is the definition of a synthetic population. The population generation framework described in Chapter 7 is used to generate the population of Ile de France including its long-distance travel demand. Therefore, it is ensured that the population meets the socio-demographics of Ile de France.

Duration and percentage-of-time targets are derived from the travel demand of the synthetic population. The activity types used are 'daily life', the standard activity type, 'business', 'vacation', 'visits' and 'private'. The

'private' activity type covers all non-business activities that are neither vacation nor visits. The 'daily life' activity is special in several aspects. First, it is the only one which is included in the activity set of every agent. Second, the destination is fixed for 'daily life', namely Ile de France. Finally, there is no duration target assigned to 'daily life'. Thus, there is no discomfort in performing 'daily life' for short or long periods.

In total 6.6 million agents with targets were generated. This is due to the fact that less than 60% of all persons undertake long-distance trips as it has been shown in Chapter 6. A monetary budget is defined via the income which is part of the synthetic population. It is assumed that 20% of the monthly income is saved. The saved money is added to an agents' monetary budget at the end of each month. The vacation day budget is dynamic as well. At the end of each month 2.5 days are added to the time budget.

8.1.2 Destinations

The set of destinations used in this application includes 27 options which are shown in Table 8.1. The granularity of the destination set decreases with distance. Destinations that are close cover small areas, while destinations that are further away cover large areas.

The destinations within France are clustered by region which is the highest administrative division in France. There are 13 regions in metropolitan France that constitute the French destination options in this application. Note that the regions used here mirror the boundaries that were introduced by the reform of metropolitan regions in 2016. Overseas regions were not considered as national destinations. On European level spatial clusters of countries are considered as destinations. Beyond the French neighbors there are only two zones that are worth to constitute a destination by itself. On the one hand, Greece attracts a substantial number of tourists. On the other hand, Scandinavian countries are visited more often for Business purposes than any other non-neighboring countries. The rest of Europe does not attract many trips and is clustered to a single destination named Eastern Europe. Outside of Europe the destinations are defined on continental level. An exception is Northern Africa which attracts substantially more travel demand than the rest of Africa which is confirmed by the ENTD 2008. The reasons are geographical proximity and a common history. Therefore, Northern Africa defines its own destination cluster.

Table 8.1: Destinations used in Ile de France (IdF) scenario

Destination	Distance to IdF [km]	Trips from IdF in ENTD [100,000]
Ile de France	0	0
Auvergne-Rhone-Alpes	424	57.9
Alpes Cote Azur	649	40.2
Bourgogne	229	30.2
Brittany	393	41.7
Centre ValdeLoire	160	45.3
Corsica	931	2.0
Grand Est	252	31.9
Hauts de France	168	38.3
Nouvelle Aquitaine	455	54.0
Normandy	168	60.8
Occitanie	588	32.0
Pays de la Loire	311	41.6
Iberia	1259	15.5
UK and Ireland	359	9.8
Italy and Malta	1000	6.2
BeNeLux	306	17.0
Germany and CH	578	6.8
Scandinavia	1563	1.3
Eastern Europe	578	6.8
Greece	2091	0.1
North Africa	1709	20.2
Rest of Africa	5341	3.3
Asia	7923	5.9
North America	7248	2.0
South America	7623	4.6
Oceania	15436	0.2

The distance between two destination clusters used here is defined by the mean distance travelled by respondents in the ENTD 2008. The set

of destination clusters and their distances from Ile de France are shown in the second column of Table 8.1. Furthermore, the weighted number of long-distance trips performed by the 2,248 respondents in the ENTD that lived in Ile de France is presented.

Table 8.1 shows that more than 80% of all long-distance trips were on national level (considering metropolitan France). The number of trips decreases substantially for destinations which are further away. For instance, Oceania attracted only three of the respondents for a long-distance trip. In fact, these trips were family visits to French Polynesia. The weights provided in the ENTD lead to an estimate of 20,000 trips to Oceania. A second case that is worth noting is South America. Again, strong historical ties play a role. A substantial amount of trips to South America end in Guadeloupe and Martinique which are both Overseas Territories of France.

8.1.3 Modes

The transport modes used in this scenario are limited to the three most frequently used modes for long-distance travel: car, train and plane. Two options are considered for the train service, the regional service TER and the French inter-city service named TGV. The main characteristic of a mode in C-TAP is the price and travel time. The travel time has two components, the travel speed and the mode access time. The access time has substantial impact on mode choice for long-distance travel. For example, accessing a plane can take several hours including travelling to the airport, checking-in, boarding, etc. The mode characteristics used in this scenario are shown in Table 8.2.

Table 8.2: Modes used in Ile de France scenario

Travel mode	Base Fare [€]	Price [€/100 km]	Speed [km/h]	Access Time [min]
Car	0	15.60	80	0
TER train	0	8.25	75	30
TGV train	0	10.45	130	60
Plane	100	7.00	600	240

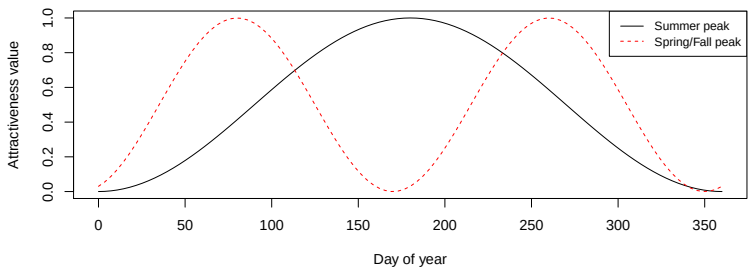
The speed of cars is close to the average that has been measured on a motorway in France using cellular phone tracking (Yim, 2003). Costs of car travel usually vary, because fuel cost and fuel consumption fluctuate. In this scenario, an assumption of 1.20€ cost per liter fuel has been made which was the average price in 2008. A consumption of 8 litres per 100 km which is the average consumption outside of urban environment was assumed. This leads to fuel costs of 9.60 € per 100 km. A motorway toll of 6€ per 100 km has to be added in the French environment. The average prices and speeds (including average delay) for the regional rail service TER and the high-speed train service TGV has been taken from a report of the French rail authorities (Arafer, 2016). An assumption of 600 km/h for airport-to-airport travel is common in recent literature (e.g. Nelldal, 1998; Schaefer and Victor, 1997). The cost structure of flight tickets includes a base-fare, but has the lowest variable cost. The cost structure is a linear approximation of log-transformed logit models that have been estimated (Brueckner and Whalen, 2000; Brueckner, 2003).

8.1.4 Seasonal Externalities

C-TAP supports several versions of seasonal attractiveness functions influencing the pull factors ϕ_λ within the decision making of the agents. First, destination attractiveness for vacations are introduced to this scenario. Three types of attractiveness functions are used: a constant function for destinations such as Asia, a summer peak function for Mediterranean destinations such as Iberia and a spring/fall peak function for North-European destinations such as United Kingdom. The latter two are illustrated in Figure 8.1. The attractiveness functions used are based on assumptions.

Business trips are not affected by yearly attraction patterns. Instead, a clear weekly pattern is evident in most data sources such as the French ENTD. Most business activities take place between Monday and Thursday and to a smaller extent on Fridays. Weekends are usually avoided for business trips. A weekly attractiveness curve mirroring this characteristic was added for business activities. The latter is independent of the considered destination. No seasonal effects on attractiveness of visits have been taken into account in this scenario. However, one can consider adding seasonal effects to model peaks on Easter and Christmas for visits.

Figure 8.1: Illustration of vacation attractiveness functions



8.2 Analysis

The application described above is used to demonstrate the validity of the concepts of C-TAP. A proof of concept has to show that C-TAP predicts long-distance travel demand as expected. Several aspects of the generated demand are compared to the expected outcome which is defined either by survey results or by synthetic values.

8.2.1 Travel Demand

The most important value that needs to be analyzed is travel demand generated by C-TAP which is a simulation of annual long-distance travel demand. Therefore, several measures have been employed to evaluate travel generated by simulating this scenario. The measures are compared to the travel demand of the synthetic population used for this scenario.

First, the total number of tours by purpose is calculated. The number of tours is shown in Table 8.3. The total numbers are comparable between the two sources. In both cases, a business trip is the least frequent long-distance trip purpose. Around 6.5 million long-distance trips are due to a business activity. Almost twice as many trips are done for vacation and private reasons. More than 16 million long-distance trips are visiting trips. Small

Table 8.3: Number of tours by purpose

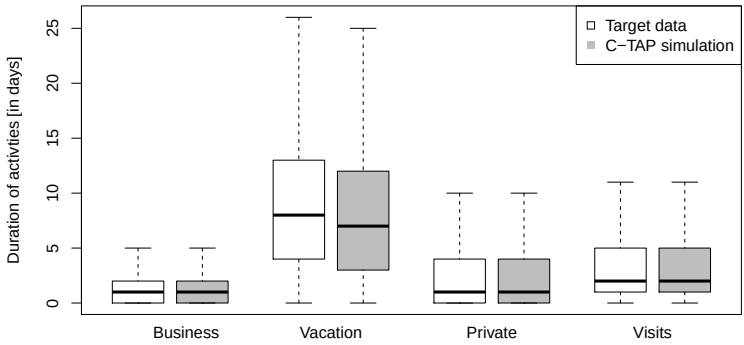
Purpose	Tours [in 1,000,000] data	simulated	Average activity duration [days] data	simulated
Business	6.43	5.9	1.80	1.77
Vacation	12.96	11.2	9.91	9.23
Private	11.72	11.3	3.13	3.18
Visits	16.52	16.1	3.98	3.89

differences are caused by random effects and can be neglected. The table also includes the average activity durations for each activity type. The average durations by purpose fit the target values very well. Vacations are typically the activities with the longest durations. All other activities usually last only a couple of days.

Second, the simulated durations of activities have to meet the durations representing the real world. Figure 8.2 illustrates the duration distribution by activity type. The box-plots of target data and simulated data are again very similar. Business activities are usually very short, namely a few days. Vacations typically last one to two weeks, but also have substantial amount of very long durations, like three or four weeks. Visits and private activities are very similar and usually last a couple of days. This is due to the fact that these activities are often tied to weekends.

In addition to the activity duration distribution the interval distribution needs to be compared. An interval is the time period between two activities of the same type. Figure 8.3 illustrates the simulated intervals by activity. The target distribution was taken from the mobile phone billing data as described in Chapter 6. Both interval statistics were limited to the same time period, namely May to October, because the mobile phone data available is limited to this period. The means and distributions of the intervals are similar for all activity types. However, the simulated interval distribution is less dispersed than the targeted mobile phone-based distribution. This can be explained by the fact that the simulated distribution is based on many more data points than the targeted distribution. Note that there is no direct control for (seasonal) frequency implemented in C-TAP. Nevertheless, the intervals recorded in the CDR data are reproduced by the C-TAP simulation.

Figure 8.2: Analysis of duration distributions by activity type



Finally, the number of long-distance trips by agent is analyzed. It is crucial that agents reproduce the number of tours given by the input data. Figure 8.4 shows the distribution of the number of tours. The two distributions are similar. Few agents or persons have only a single long-distance tour per year. However, more than 35% of persons have two or three tours per year. This is replicated well by the simulation. A second peak appears at six yearly tours and the tail of the distributions is flat for both cases.

Long-distance travel demand is the core of the simulation output. Therefore, it is essential that the demand simulated meets the target values which describe the real world. Table 8.3 and Figures 8.2, 8.3 and 8.4 have illustrated in this section that the demand simulated is reasonable and mirrors the targeted values.

Figure 8.3: Analysis of interval distributions by activity type

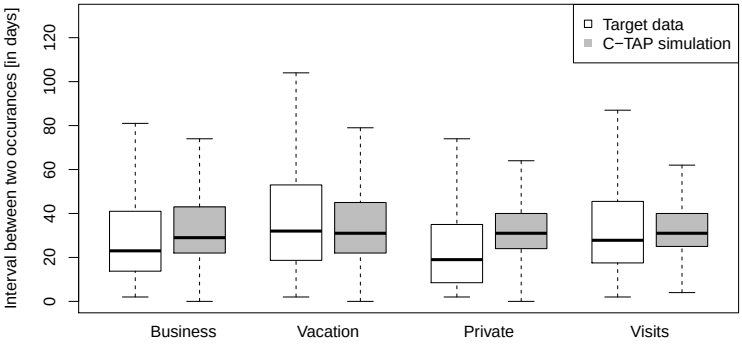
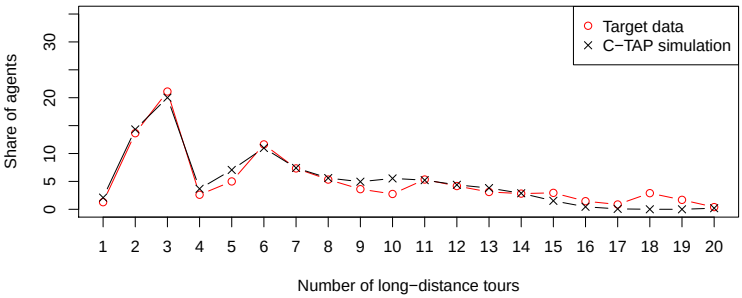


Figure 8.4: Distribution of number of tours by agent



8.2.2 Destination Choice

Furthermore, destination choice is a major part of the C-TAP implementation. It is crucial that agents travel to the same destinations as it is observed in reality. As described above, the underlying scenario was initialized with 27 destinations covering the whole world. The target shares of the attracted trips to these destinations were taken from the ENTD 2018 (see Section 6.2 for a description of the data source). Table 8.4 (national travel) and Table 8.5 (international travel) show the comparison between the target shares and the simulated shares by purpose. Trips to the home destination (Ile de France) have been neglected since they dominate the number of total tours.

The results show that the share of attracted trips fits the target shares. Each trip purpose reflects the given destination shares. As before, the major part of the trips had a national destination. Most business trips are attracted by the region of Auvergene-Rhone-Alpes which is home of the metropolitan region of Lyon. Vacations tend to take place on the coastal areas such as Alpes-Cote-Azur or Occitanie. Visits and private activities are more diffused than the other two activity types. Focusing on international travel, Table 8.4 shows that most business activities take place in the proximity, namely in Germany, the United Kingdom and the Benelux states. This does not hold for vacations. Same as on national level, coastal regions are preferred, e.g. Iberia or Northern Africa.

8.2.3 Seasonal Influence

Seasonal effects have an impact on long-distance travel behavior. C-TAP takes into account several seasonal effects. It has to be illustrated that these effects are mirrored in the simulation output, namely the travel demand.

First, destination attractiveness plays a crucial role for vacation planning. European destinations have been attached with two different seasonal attractiveness functions for vacation trips. Alpes-Cote-Azur is picked for further analysis and its attracted number of vacation trips throughout the simulated year is illustrated in Figure 8.5. Before the 8th week and after the 42nd week vacations are very rare at this destination. This is due to the low attractiveness at this time of the year. On the other hand, the number of vacations peak during the summer months. The share of attracted vacations is the highest between the 20th and 30th week. This is the time of the year where attractiveness of summer-peak destinations is superior to destinations

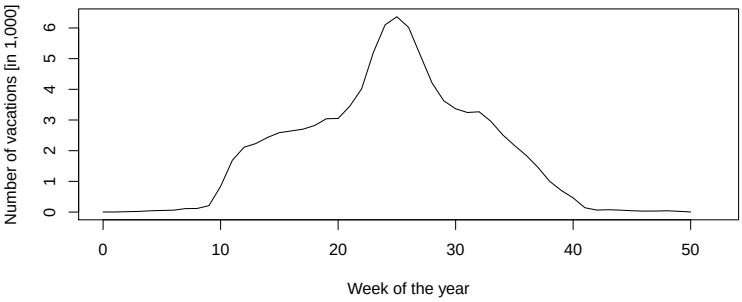
Table 8.4: National destinations accessed in Ile de France scenario

Destination	Business		Vacation		Visits		Private	
	Target	C-TAP	Target	C-TAP	Target	C-TAP	Target	C-TAP
Auvergne-Rhone-Alpes	16.4%	16.3%	11.5%	11.9%	13.3%	15.6%	14.0%	14.0%
Alpes Cote Azur	4.3%	3.4%	12.1%	12.4%	5.7%	5.4%	8.2%	8.9%
Bourgogne	3.8%	3.1%	3.2%	4.5%	6.3%	6.3%	4.9%	4.5%
Brittany	5.4%	5.8%	7.5%	8.0%	6.6%	6.8%	5.7%	5.4%
Centre Val de Loire	5.0%	5.0%	2.8%	4.5%	5.8%	5.2%	3.2%	2.3%
Corsica	0.0%	0.0%	1.2%	0.6%	0.2%	0.2%	0.0%	0.0%
Grand Est	4.8%	4.7%	3.6%	3.2%	8.4%	9.2%	5.4%	4.8%
Hauts de France	10.0%	11.2%	2.3%	1.2%	6.9%	6.7%	6.4%	6.0%
Nouvelle Aquitaine	12.2%	12.8%	10.7%	10.4%	14.0%	13.5%	11.9%	12.7%
Normandy	3.7%	4.2%	3.7%	2.7%	7.3%	7.3%	7.3%	8.4%
Occitanie	9.7%	9.5%	12.7%	13.3%	11.6%	10.8%	12.4%	12.6%
Pays de la Loire	9.4%	9.8%	4.9%	4.7%	8.0%	8.6%	6.3%	7.4%

Table 8.5: International destinations accessed in Ile de France scenario

	Business		Vacation		Visits		Private	
Destination	Target	C-TAP	Target	C-TAP	Target	C-TAP	Target	C-TAP
Iberia	0.6%	0.0%	5.1%	5.5%	1.0%	0.5%	3.0%	3.6%
UK and Ireland	1.7%	0.5%	1.1%	0.9%	0.4%	0.0%	1.1%	0.2%
Italy and Malta	1.3%	1.2%	2.7%	2.3%	0.7%	0.6%	0.7%	0.0%
BeNeLux	5.2%	6.1%	1.3%	1.7%	1.1%	1.4%	3.5%	4.6%
Germany and CH	2.4%	3.1%	1.6%	2.4%	0.4%	0.0%	3.1%	3.4%
Scandinavia	0.2%	0.0%	0.3%	0.1%	0.0%	0.1%	0.0%	0.0%
Eastern Europe	1.5%	1.4%	1.8%	1.9%	0.0%	0.0%	0.5%	0.0%
Greece	0.0%	0.0%	1.0%	1.1%	0.1%	0.1%	0.1%	0.0%
North Africa	0.3%	0.0%	3.7%	2.4%	1.5%	1.1%	0.9%	0.8%
Rest of Africa	0.1%	0.0%	0.9%	0.2%	0.2%	0.1%	0.1%	0.0%
North America	0.5%	0.4%	0.4%	0.1%	0.1%	0.1%	0.5%	0.5%
South America	0.4%	0.5%	1.0%	2.1%	0.2%	0.1%	0.2%	0.0%
Oceania	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%	0.1%
Asia	0.9%	1.1%	0.8%	1.9%	0.2%	0.1%	0.2%	0.0%

Figure 8.5: Attracted number of vacations for Alpes-Cote-Azur



with a spring/fall peak (see Figure 8.1). The curve for the number of activities is a good fit to the curve describing the seasonal attractiveness. Thus, seasonal effects can be directly included in the simulations.

Furthermore, day-of-week attractiveness has been introduced to business activities. The number of all business activities by day-of-week are shown in Table 8.6. It can be noticed that most business activities take place outside of the weekend. Nevertheless, there are still some business activities on the weekends. A substantial share of these activities are business activities with long durations (more than 10 days). Weekends can not be avoided for these activities.

Table 8.6: Number of business activities by day-of-week [in 1,000,000]

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Business activities	3.1	4.2	4.7	4.6	3.7	2.6	2.2

8.2.4 Mode Choice

Mode choice is important to the value of C-TAP for applicants of the model since long-distance modal split is usually targeted by infrastructural changes. Section 10.1 gives two examples for possible fields of application. The modal split is highly dependent on the destination that is considered. More precisely, the distance travelled is the main variable in the modal choice.

Every destination outside of Europe as well as Greece, Scandinavia and Eastern Europe was accessed by plane on every trip. Very short trips to neighboring destinations were undertaken by car or by regional train if a car is not available to the corresponding agent. Both results are reasonable and are confirmed by the ENTD survey.

Table 8.7 shows the mode share for all trips between Ile de France and Alpes-Cote-Azur. These two destinations are chosen for analysis since all modes are competitive for trips between these two regions. In particular, the travel time difference between a TGV train and the airplane is less than an hour (including access time). The travel times in the scenario match the average travel times reported in the ENTD. This confirms that the assumptions made in Table 8.2 are reasonable.

One can see that the slow and expensive option, namely the car, is rarely used since the overall travel time is not sufficient despite the low access time. Most of the agents chose the TGV which is almost as fast as the fastest mode, namely the airplane. The TGV is substantially cheaper and ,thus, it is a reasonable trade-off of costs and travel time. Agents that are not cost sensitive tend to chose the airplane which is the fastest and most expensive option. These agents have either high income or were travelling for a business activity.

Table 8.7 shows also the modal split reported in the ENTD. Coaches and motorcycles were also chosen by the respondents, but these modes have a negligible total share of less than 0.2%. Car travel is underestimated in this scenario due to the fact that the car is never the cheapest option for long distances. However, unlike for other modes car costs do not increase when travelling with multiple persons from a household. The respondents of the ENTD reported that there were more than two persons in the car for 86% of all car trips between the two regions. This fact demonstrates that joint travel decisions are crucial for a simulation of long-distance travel demand. Therefore, a household structure needs to be implemented in C-TAP which

is one of the future challenges. In addition, access time has been fixed in this scenario. Heterogeneity for access times needs to be added to further improve the model.

Table 8.7: Modal Split for Trips between Ile de France and Alpes-Cote-Azur

Mode	Travel time [h:min]	Costs [€]	Modal Split C-TAP	Modal Split ENTD
Car	8:07	101.24	8%	42%
TER train	9:09	53.54	11%	3%
TGV train	5:59	67.82	49%	36%
Plane	5:05	145.43	32%	19%

8.3 Conclusion

An application of C-TAP has been presented, namely the Ile de France scenario. A simulation of the scenario estimated the long-distance travel demand for more than 6.5 million agents. C-TAP reproduced the given travel demand very well. In addition, destination attractiveness and seasonal effects meet the targeted control values.

Some limitations of the approach have been pointed out. First, household structures are necessary and important for analysis of long-distance travel demand. Households are missing in the current implementation of C-TAP, but are potentially of interest for further development. Second, mode choice is based on objective parameters, namely travel time and monetary costs. Introducing individual preferences for modes may further improve the model.

Overall, this chapter demonstrated that C-TAP generates long-distance travel demand that mirrors existing data sources. The following chapter will add value to an implementation of a continuous simulation by illustrating the efficiency of such a simulation.

Chapter 9

Implementation Details and Operational Challenges

There are operational challenges and implementation details that have not been discussed yet. The theoretical advantages of C-TAP as a continuous target-based simulation have been presented in the previous chapters. However, it remains to be shown that efficiency can be achieved in practice which is important to prove the value of continuous activity planning for simulating long-distance travel demand. Before focusing on details of the implementation, it is worth noting that the software used in this work has been implemented in the C++ language.

The remainder of this chapter is structured as follows. First, the user control techniques and parameters are introduced. Afterwards, it is shown how these parameters can be used for calibration. Finally, the efficiency of C-TAP is illustrated with some example simulations.

9.1 Interfaces

One of the main advantages of agent-based simulations is flexibility. Each agent can be initialized with different parameters yielding variety in behavior on the microscopic level. The drawback that comes along with this advantage is the initialization process of the simulation, which is crucial and requires additional effort.

The different control mechanisms and interfaces of C-TAP needed to initialize the simulation are summed up in the following. First, global parameters control the simulation output and influence simulation run time. For instance, the number of promising alternatives considered during

the activity planning effects the run time substantially. Second, the main simulation input is a synthetic population. A population needed for C-TAP has two aspects: It has to meet distributions and marginal totals of socio-demographics. In addition, the travel demand which is transformed to behavioral targets of the agents has to meet reality. Furthermore, a network including destinations and their attractiveness needs to be provided. The modes available to the agents have to be defined. Finally, seasonal effects on attractiveness, travel times and prices have to be initialized.

The drawback of the flexibility C-TAP offers is the amount of data needed to initialize the simulation. Similar to most agent-based approaches users of the simulation need to collect substantial amount of data before running the simulation. First, data is needed to generate a synthetic population with a reasonable level of detail (see Chapter 7). Second, data is needed to calibrate the simulation. Calibration is crucial for the applicability of a travel demand simulation and is discussed in the following section.

9.2 Calibration and Validation

Any agent-based model is in need of calibration in order to initialize all parameters of the model with meaningful values. Given a set of parameters which define the model and a set of control variables which are calculated by the model, it is necessary to find the optimal values for the model parameters. A set of parameters is optimal if they yield minimal distance between simulated and observed control variables. Control variables in a travel demand simulation might be counts of specific modes or links which are collected. Due to the complexity of the problem an estimate of the optimal solution is usually sufficient.

Several algorithms solving this problem have been presented in recent history. Most of the approaches build on the Simultaneous Perturbation Stochastic Approximation (SPSA) method introduced by Spall (1992). SPSA is an iterative approach. In each iteration variations of model parameters are considered. Following specific rules Spall has proven convergence for his algorithm. SPSA has been further improved, e.g. Lu *et al.* (2015) introduce a weighted version of SPSA improving the performance of SPSA. Flötteröd (2017) accelerated the search and applied his framework to transport simulations, e.g. MATSim (Agarwal *et al.*,

2017).

There are several parameters needed to be calibrated in C-TAP. These parameters are mainly destination-based, e.g. destination attractiveness. A detailed calibration was not performed since calibration needs additional data such as counts of airport users or train tickets. Information on travelled distances has large value for this purpose. However, a reliable source for this data is hard to find and is a challenge for future applications of the simulation.

A second step that was not within the scope of this thesis is validation of the simulation. Validation is needed to prove the applicability of the simulation. Two ideas can be implemented to validate an agent-based simulation for travel demand. The first approach takes statistics that describe long-distance travel but were not used for calibration, e.g. counts of mode usages. Second, one can use a calibrated scenario and introduce infrastructural changes that happened in reality. Agents in C-TAP ideally adapt their travel behavior the same way the real world population has done. However, C-TAP has not been validated yet, because validation requires further data.

9.3 Scalability and Efficiency

One of the main advantages of continuous activity planning is the scalability. Theoretically, the memory used as well as the simulation run time scale linearly with the problem size. The scalability is illustrated in the following. A reference scenario is compared to simple modifications of the scenario. The reference case used for comparisons in the remainder of this section is a sub-scenario of the Ile de France application used in Chapter 8. The reference scenario uses the same parameters and destination set, but the population is limited to the residents of the city of Paris.

All simulation runs presented in the remainder of this section were performed on the same machine. The machine runs at 3.00 GHz (x86_64 Architecture) and has 384 GB of RAM available.

To begin with, the population size and its impact on the simulation is analyzed. Long-distance travel demand of three alternative populations in the same scenario were simulated. First, a random one-percent sample of the Ile de France population is simulated. Second, the whole Ile de France

population as described in Chapter 8 is simulated. Third, travel demand for the whole population of France is simulated. The comparison of the three simulation runs and the reference scenario is shown in Table 9.1.

Table 9.1: Impact of population size on runtime and memory

	Population	Runtime [h:min] total	Runtime [h:min] per $10e^5$ agents	Memory [RAM] total
1% Ile de France	73,687	0:14	0:19	0.58 GB
Reference (Paris)	1,384,216	3:14	0:14	9.95 GB
Ile de France	6,665,460	21:49	0:20	56.61 GB
France	34,880,982	224:23	0:21	295.81 GB

Table 9.1 shows that population is the main memory consumer in this scenario definition. A change in population size directly impacts the memory needed for the simulation. However, the memory a single agent consumes is not constant, but, among other factors, dependent on the number of targets and activities. The simulation runtime is effected by the population size as well. To be more precise, the number of decisions that have to be calculated is driven by the population size. Each decision calculated includes construction of all feasible alternatives and optimization of promising alternatives. All decisions constitute the activity planning and are the main component of the simulation runtime.

Furthermore, it is fundamental that a continuous simulation such as C-TAP simulates arbitrarily long periods of time. The reference scenario covers one year of long-distance travel demand. Table 9.2 shows the impact of an increase in simulated time. The simulated time was increased to two years as well as to three years of long-distance travel demand.

Table 9.2: Impact of simulated time on runtime and memory

	Simulated time	Runtime [h:min]	Memory [RAM]
Reference	1 year	3:14	9.95 GB
	2 year	5:12	9.95 GB
	3 year	8:34	9.95 GB

One can see that the simulated time has no effect on memory needed for the simulation. The data describing the current state of the simulation is stable and independent of the time simulated which is one of the advantages of a continuous simulation. Neither a decision history nor an activity plan has to be maintained during the simulation. All decisions are based on targets and (dynamic) state values. The simulation runtime grows linearly with the simulated time which is the second advantage of continuous planning.

In addition, the network size is potentially important for the simulation runtime. The number of destinations defines the number of alternatives for each agent. Each promising alternative has to be optimized, i.e. the optimal sequence of durations has to be the calculated. Table 9.3 shows experimental simulation runs where the network size of the reference scenario has been extended.

Table 9.3: Impact of network size on runtime and memory

	Destinations	Runtime [h:min]	Memory (RAM)
Reference	27	3:14	9.95 GB
Scenario 1	53	4:28	9.95 GB
Scenario 2	79	5:11	9.95 GB

Two alternative scenarios were considered. First, the reference network has been copied twice, i.e. each destination of the reference network has two copies in the alternative scenario. Second, the reference network has been copied three times. One can see that the impact of the network size is limited. The memory needed does not change substantially. This is due to the fact that the network structure does not consume a lot of memory. Therefore, memory is not a relevant issue if the network size has the given magnitude. However, simulation runtime increases with a growing network. An agent has to consider all destinations for his set of feasible alternatives. Nevertheless, the effect is sub-linear, because the number of alternatives which are optimized does not increase. It remains equal to the number of promising alternatives kept. The number of promising alternatives in the scenarios presented in Table 9.3 is equal to one per activity. The effect of this parameter is larger than the network size as it has been shown in

Section 5.4.2. However, it is worth noting that an increase of network size adds complexity to the calibration process

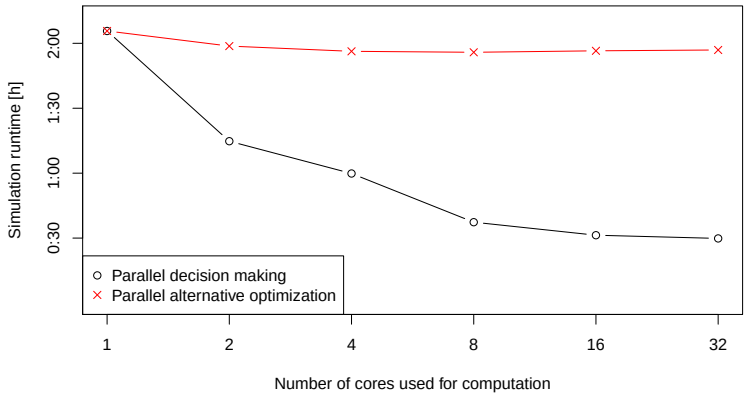
9.3.1 Parallelization

The analysis above shows that C-TAP scales very well with the problem size. This characteristic is crucial in order to run a simulation on large scales in terms of population, simulated time and network. However, as for most simulations there is further potential to improve the simulation runtime, namely parallelization.

First, parallelization was implemented within the main simulation loop (see Algorithm 3.1). Each time an agent becomes idle it needs to make a decision on his next activity. The decision is calculated within the activity planning module. Several planning modules can run in parallel, i.e. a module is generated for each parallel thread. Therefore, several agents can calculate their decisions simultaneously. For example, running C-TAP on 8 parallel threads gives the opportunity to calculate 8 decisions in parallel. However, C-TAP is an event-based simulation. As a consequence it is not guaranteed that agents become idle at the same time. An agent might be assigned to a planning module before all agents that became idle earlier have calculated their decision. Nevertheless, when simulating large populations as described in Chapter 8 the difference is relatively low. For example, running the scenario described in Chapter 8 with 32 threads leads to the following situation: The maximal time difference of two agents that are calculating their decisions in parallel is less than one hour of simulated time. In other words, the worst case is that an agent makes his decision based on information which is one hour old. This parallelization option has to be neglected, if a user has to ensure that all events and calculations are performed in chronological order.

A second piece of the simulation is parallelized as well. On the one hand, it ensures chronological calculations. On the other hand, the speed up is smaller and has a boundary. The calculation within the activity planning module can be run in parallel as well. As described in Chapter 5 an optimal set of durations has to be calculated for each feasible alternative. An alternative is a sequence of activities, destinations and modes. The alternatives are independent and, thus, can be optimized in parallel using the Downhill-Simplex algorithm (see Section 4.3). This technique has less

Figure 9.1: Runtimes of parallel computing approach for the Paris scenario



potential speed-up, because some calculations are still sequential, e.g. the generation of the alternatives. The potential of this approach is bounded by the number of alternatives that are optimized in the activity planning.

Parallelization of C-TAP has been implemented using the OpenMP library (Dagum and Menon, 1998) which is a standard API for shared-memory programming using C++ and other programming languages. C-TAP simulations presented in this section used the latest version OpenMP v4.5. These simulation were executed on a scientific compute cluster of ETH Zurich (*Euler,* ETH, 2019) which has several thousand computing nodes and is suited for parallel calculations.

The impact of both parallelization techniques is analyzed. The scenario used for evaluation is the reference scenario of the Paris population used earlier in this section. The runtime results of these simulation executions are shown in Figure 9.1. The simulation has been executed using 1,2,4,8,16 and 32 threads. One can see that substantial speed-up can be achieved by using the first paralleling option where several activity planning modules run in parallel. However, saturation is achieved with 16 parallel threads.

Adding more cores does not improve the efficiency of the simulation, because the overhead to spawn new threads exceeds the speed-up of parallel computation. Furthermore, a single priority queue has to be maintained by the main thread. This means that only a single thread can perform operations on the priority queue at a single point of time. Any time the priority queue is accessed by a thread all other idle threads have to wait to gain access. The second approach does not have any impact on this scenario. The main reason is the low number of promising alternatives (one per activity type) that is kept for optimizing. Using 32 threads increases the runtime in comparison to 16 threads. As before, this is due to the overhead of spawning new threads.

9.4 Conclusion

This chapter has pointed out the main drawbacks and advantages of the simulation approach presented in this work. An advantage is the high flexibility and heterogeneity which is a characteristic of agent-based simulations. This characteristic constitutes the main disadvantage. Much data is needed to feed all interfaces of the model. This data is hard to gather, in particular for the case of long-distance travel. In addition, calibration effort is needed. However, a further advantage justifies the implementation of the target-based framework for long-distance travel demand simulations. That is the efficiency of the continuous target-based planning. It has been shown that the framework scales very well in any dimension in terms of simulation runtime and memory needed. Scalability of the framework is supported by the potential to parallelize the approach and further improve runtime. Therefore, C-TAP is predestined to be used for large-scale simulations of travel demand.

Chapter 10

Discussion and Outlook

10.1 Field of Application

The main field of application for a simulation of long-distance travel demand is cost-benefit analysis of substantial infrastructural changes or policy adjustments. Simulation results may support policy makers when considering infrastructural investments. Two potential applications which both have a recent real-world example are discussed in the following.

The first application is construction of new tunnels for rail or street traffic, e.g. the new Gotthard rail tunnel opened in 2016 (Swiss Railways, 2016). A new tunnel impacts the activity planning via mode specific parameters. The travel time for some destinations typically decreases substantially yielding more agents taking this alternative. Similar effects apply to new high-speed train connections.

A second field of application is the construction of a new airport such as the one in Berlin (Bubalo and Daduna, 2011). Same as in the aforementioned example new airports are usually planned to replace older versions. Therefore, this infrastructural change does not introduce a new mode to the simulation. However, other impacts are modeled by C-TAP. This includes decrease of travel costs, potentially change in access time due to better accessibility, and, most important, decrease in travel time and cost to distant destinations. For instance, a new airport can decrease travel times to distant areas, because the number of transfers needed is reduced.

10.2 Outlook

This thesis illustrated that the implementation of C-TAP can be a valuable tool to analyze and predict long-distance travel demand. However, there is potential for further improvement. This section gives an overview of potential future steps that are needed to increase the value of the model.

The first promising improvement of the simulation presented in the previous chapters is the data used to initialize and calibrate the model. It describes the state of 2007 and, therefore, is outdated. In addition, the CDR data used has certain inaccuracies as described in Chapter 6. Therefore, more recent and more accurate individual travel data would be of great interest to synthesize a realistic population. The same request is also valid for mode specific and destination specific data. The information utilized in this thesis is based on a national travel survey. However, information on the road network and airport as well as train schedules are crucial to generate a multi-modal network. Hence, a focus for future usage of the model presented is data collection.

More and recent data improving the quality of the population, the network or the destination description is valuable. Nevertheless, collection of additional data is also crucial for an empirical validation of the model which is needed to illustrate the potential value of C-TAP. Empirical validation can be performed on two different levels. In both cases the model has to be initialized and calibrated appropriately. The first idea is to find independent data sources that are not used for calibration, e.g. counts of certain roads. This data has to be compared to the output generated by the simulation. However, a major concern here is the inconsistency of independent data sources. The second validation approach is a sensitivity analysis in two steps. First, the simulated scenarios are mutated and a hypothesis on the triggered simulation effects is stated. After evaluation of the hypothesis, the approach is iterated through all potential scenario mutations. Second, input-output transformation of a real world example is monitored. In other words, a scenario with a substantial change of the infrastructure observed in the real world is chosen. Afterwards, it is evaluated whether the scenario modifications yield a transformation of the model output that matches the real world impacts.

Detailed initialization as well as validation improves the value of the model as presented in this thesis. However, further work on the model itself

may advance the simulation. The major tasks for further development have been mentioned in this work at the relevant points and are summarized below. Limitations addressed by future development of C-TAP include the modelling of agents. First, family origin of agents has to be included in the definition of an agent since in is an important driver for the destination choice of private activities. This information is rarely gathered in existing (long-distance) travel surveys, but a data source providing this information can improve C-TAP results. Second, a household structure needs to be introduced since it has a non-negligible effect on long-distance travel. For instance, mode choice and destination choice for vacations is affected by the household structure. Third, impact of the human day rhythm has to be considered, i.e. the need for sleep limits the time that an agent can spend in a car.

Furthermore, implementation details have room for improvement, e.g. the Nelder-Mead approach can be potentially replaced by a method using the derivative to iterate towards a local optimum such as the steepest descent method (Fletcher and Powell, 1963). In addition, further speed-up of the simulation runtime can be potentially achieved by exploring further parallelization techniques.

Finally, the output of an updated C-TAP model can be combined with models of daily life to get a complete picture of travel demand. Two types of daily life models can potentially enrich C-TAP simulation output. First, results of a local microsimulation such as MATSim (Horni *et al.*, 2016) can be used to generate the total travel demand of a spatially bounded population. Second, a macro-model such as a model of speed and volume (e.g. Sarlas and Axhausen, 2015, 2016) can be used to improve mode and destination choice within C-TAP.

Despite the limitations and potential future work described above, the model presented in this thesis is an important contribution to the field of travel demand modelling as it is pointed out in the next section

10.3 Contribution

This work closes an existing research gap. A tool to simulate annual long-distance travel demand has been presented here. So far such a tool has not been available to researchers and policy makers. This is a valuable

contribution since the share of long-distance travel is growing. Therefore, the need for analysis and prediction tools is rising as well.

The concept of continuous target-based planning has been utilized to implement an agent-based simulation of long-distance travel demand. This concept and its predecessor, the need-based theory, have previously not been used for large scale agent-based simulations. Furthermore, flexible destination choice has not been a concept implemented in agent-based simulations. Finally, data collection using alternative data sources such as big data are rarely used to synthesize populations due to missing semantic information. The population synthesis presented in this work overcomes this issue by combining big data sources with traditional surveys to get reliable information. In general, despite the need of reliable prediction tools an agent-based simulation of long-distance travel demand has not been proposed before. This gap is closed by the underlying work.

10.4 Conclusions

The goal of this thesis has been the development of an agent-based simulation for long-distance travel demand. The idea of continuous target-based activity planning has been utilized for this task. The advantage of a continuous approach is the flexibility in planning. Neither the sequence nor the schedule of activities is fixed. Both decisions are made during the simulation. This characteristic is suitable for a simulation of long-distance travel demand.

As for any other agent-based simulation, one of the major difficulties is collection of data needed for initialization and calibration. The focus on long-distance travel leads to a challenging task since long-distance surveys are rare and unreliable. This work demonstrated that big data can resolve this problem. Modern data sources such as mobile phone data provide valuable information that can be used to generate synthetic populations.

An application to a real-world scenario has shown that current travel demand can be mirrored by C-TAP. The simulated demand is not limited to activity type and duration planning, but also includes reasonable destination and mode choice. It has been demonstrated that efficient implementation is possible. Thus, it is easy to scale scenarios in any dimension such as population size or simulated time. In addition, parallelization further improves the usability of the model.

Future work is needed in two different areas: On the one hand the simulation needs additional functionality such as household structures. On the other hand, further data collection is necessary to improve the explanatory power of the model. Nevertheless, a C-TAP implementation as presented in this work is a solid foundation for future research. It is the first implementation of an agent-based simulation of long-distance travel demand and a valuable tool for analysts and policy makers.

References

Adler, T. J. and M. E. Ben-Akiva (1976) Joint-choice model for frequency, destination and travel mode for shopping trips, *Transportation Research Record*, **569**, 136–150.

Agarwal, A., G. Flötteröd and K. Nagel (2017) Calibration of behavioral parameters in an agent-based transport simulation, paper presented at the *6th Symposium of the European Association for Research in Transportation*, Haifa, September 2017.

Ahas, R., A. Aasa, Ü. Mark, T. Pae and A. Kull (2007) Seasonal tourism spaces in Estonia: Case study with mobile positioning data, *Tourism Management*, **28** (3) 898–910.

Ahas, R., A. Aasa, A. Roose, Ü. Mark and S. Silm (2008a) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study, *Tourism Management*, **29** (3) 469–486.

Ahas, R., S. Silm, O. Järv, E. Saluveer and M. Tiru (2010) Using mobile positioning data to model locations meaningful to users of mobile phones, *Journal of Urban Technology*, **17** (1) 3–27.

Ahas, R., S. Silm, E. Saluveer and O. Järv (2008b) Modelling home and work locations of populations using passive mobile positioning data, in G. Gartner and K. Rehrl (eds.) *Location Based Services and TeleCartography II: From Sensor Fusion to Context Models*, 1st edn., 301–315, Springer.

Anas, A. (1981) The estimation of multinomial logit models of joint location and travel mode choice from aggregated data, *Journal of Regional Science*, **21** (2) 223–242.

Anda, C., A. L. Erath and P. J. Fourie (2017) Transport modelling in the age of big data, *The International Journal of Urban Sciences*, **21** (1) 19–42.

Anda, C. and S. A. Ordóñez Medina (2017) Extending the hidden Markov model for activity scheduling, paper presented at the *6th Symposium of the European Association for Research in Transportation*, Haifa, September 2017.

Arafer (2016) The French passenger rail transport market 2015-2016, *Technical Report*, Autorité de Régulation des Activités Ferroviaires (Arafer), Le Mans.

ARCEP (2016) Le Suivi des Indicateurs Mobiles - les chiffres au 31 décembre 2007 (publication le 4 février 2008), webpage, March 2016, URL `http://www.arcep.fr/index.php?id=9545`.

Arentze, T. A., T. A. Pelizaro and H. J. P. Timmermans (2010) An agent-based micro-simulation framework for modelling of dynamic activity-travel rescheduling decisions, *International Journal of Geographical Information Science*, **24** (8) 1149–1170.

Arentze, T. A. and H. J. P. Timmermans (2004) Albatross: A Learning-Based Transportation Oriented Simulation System, *Transportation Research Part B: Methodological*, **38** (7) 613–633.

Arentze, T. A. and H. J. P. Timmermans (2006) A new theory of dynamic activity generation, paper presented at the *85th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2006.

Arentze, T. A. and H. J. P. Timmermans (2009) A need-based model of multi-day, multi-person activity generation, *Transportation Research Part B: Methodological*, **43** (2) 251–265.

Arentze, T. A., H. J. P. Timmermans, F. Hofman and N. Kalfs (2000) Data needs, data collection, and data quality requirements of activity-based transport demand models, *Transportation Research Circular*, **1** (E-C008) 30–60.

Armoogum, J., J.-P. Hubert, D. Francois, B. Roumier, M. Robin and S. Roux (2008) Enquête nationale transport et déplacements 2007-2008, *Technical Report*, IFSTTAR, Paris.

Armoogum, J. and J.-L. Madre (2002) Weighting and correcting long-distance travel surveys, in K. W. Axhausen, J.-L. Madre, J. W. Polak and P. L. Toint (eds.) *Capturing Long-Distance Travel*, 152–170, Research Studies Press, Hertfordshire.

Ashley, D., J. Richardson and S. E. Young (2009) Recent information on the under-reporting of trips in household travel surveys, paper presented at the *Australasian Transport Research Forum 2011*, Auckland, September 2009.

Axhausen, K. W. and T. Gärling (1992) Activity based approaches to travel analysis: Conceptual frameworks, models and research problems, *Transport Reviews*, **12** (4) 323–341.

Axhausen, K. W. and R. Herz (1989) Simulating activity chains: German approach, *Journal of Transportation Engineering*, **115** (3) 316–325.

Axhausen, K. W., J.-L. Madre, J. W. Polak and P. L. Toint (eds.) (2002a) *Capturing Long-Distance Travel*, Research Studies Press, Hertfordshire.

Axhausen, K. W., B. Schmid and C. Weis (2015) Predicting response rates updated, *Arbeitsberichte Verkehrs- und Raumplanung*, **1063**, IVT, ETH Zurich, Zurich.

Axhausen, K. W. and C. Weis (2010) Predicting response rate: A natural experiment, *Survey Practice*, **3** (2).

Axhausen, K. W. and M. Youssefzadeh (1999) MEST - Methods for European Surveys of Travel Behaviour, *Final Report*, University of Innsbruck, Innsbruck.

Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser and T. Haupt (2002b) Observing the rhythms of daily life: A six-week travel diary, *Transportation*, **29** (2) 95–124.

Balmer, M. (2007) Travel demand modeling for multi-agent traffic simulations: Algorithms and systems, Ph.D. Thesis, ETH Zurich, Zurich, May 2007.

Balmer, M., K. W. Axhausen and K. Nagel (2006) An agent-based demand-modeling framework for large scale micro-simulations, *Transportation Research Record*, **1985**, 125–134.

Bar-Gera, H., K. Konduri, B. Sana, X. Ye and R. M. Pendyala (2009) Estimating survey weights with multiple constraints using entropy optimization methods, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

Bekhor, S., Y. Cohen and C. Solomon (2013) Evaluating long-distance travel patterns in Israel by tracking cellular phone positions, *Journal of Advanced Transportation*, **47** (4) 435–446.

Bekhor, S. and J. N. Prashker (2008) GEV-based destination choice models that account for unobserved similarities among alternatives, *Transportation Research Part B: Methodological*, **42** (3) 243–262.

Beser, M. and S. Algers (2001) Sampers - the new swedish national travel demand forecasting tool, in L. Lundqvist and L.-G. Mattsson (eds.) *National Transport Models: Recent Developments and Prospects*, Advances in Spatial Science, 101–118, Springer, Stockholm.

Bhat, C. R. (1995) A heteroscedastic extreme value model of intercity travel mode choice, *Transportation Research Part B*, **29** (6) 471–483.

Bhat, C. R. (1996) A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity, *Transportation Research Part B*, **30** (3) 189–207.

Bhat, C. R., S. Astroza and A. C. Bhat (2016) On allowing a general form for unobserved heterogeneity in the multiple discrete–continuous probit model: Formulation and application to tourism travel, *Transportation Research Part B*, **86**, 223–249.

Bhat, C. R., S. Srinivasan, J. Y. Guo and A. Sivakumar (2003) Activity-based travel-demand modeling for metropolitan areas in Texas: A micro-simulation framework for forecasting, *Final Report*, Texas Department of Transportation, University of Texas, Austin, February 2003.

Bieger, T. and C. Lässer (2008) Travel market Switzerland 2007, *Research Report*, Schweizerische Verkehrswissenschaftliche Gesellschaft (SVWG), St. Gallen.

Bierce, E. and D. L. Kurth (2014) The use of three surveys for long distance travel estimates in california, paper presented at the *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2014.

Blondel, V. D., A. Decuyper and G. Krings (2015) A survey of results on mobile phone datasets analysis, *The European Physical Journal - Data Science*, **4** (1), February 2015.

BMVI (2012) Oesterreich Unterwegs, webpage, URL `http://www.oesterreich-unterwegs.at`. Accessed on 26/04/2016.

Bonabeau, E. (2002) Agent-based modeling: Methods and techniques for simulating human systems, *Proceedings of the National Academy of Sciences of the United States of America*, **99** (3) 7280–7287.

Bonnel, P., E. Hombourger, A.-M. Olteanu-Raimond and Z. Smoreda (2015) Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations, *Transportation Research Procedia*, **11**, 381–398.

Bonnel, P., J. Zmud, M. E. H. Lee-Gosselin and J.-L. Madre (eds.) (2009) *Transport Survey Methods: Keeping Up with a Changing World*, Emerald, Bingley.

Bradburn, N. M., L. J. Rips and S. K. Shevell (1987) Answering autobiographical questions: the impact of memory and inference on surveys, *Science*, **236** (4798) 157–161.

Bradley, M. A., M. Outwater and B. Gudzinas (2015) A National Level, Tour-Based Model System to Forecast Long Distance Passenger Travel in the U.S., paper presented at the *14th International Conference on Travel Behaviour Research (IATBR)*, Windsor, July 2015.

Bricka, S. (1999) Variations in Long-Distance Travel, *Transportation Research Circular*, **1** (E-C026) 197–206.

Bricka, S. and C. R. Bhat (2006) A comparative analysis of GPS-based and travel survey-based data, *Transportation Research Record*, **1972**, 9–20.

Brög, W., E. Erl, G. Sammer and B. Schulze (2003) DATELINE – Design and Application of a Travel Survey for Long-distance Trips Based on an International Network of Expertise – Concept and Methodology

Household Planning of Car Use: Implementation of Prospective Car Logs, paper presented at the *10th International Conference on Travel Behaviour Research (IATBR)*, Lucerne, August 2003.

Brueckner, J. K. (2003) International Airfares in the Age of Alliances: The Effects of Codesharing and Antitrust Immunity, *Review of Economics and Statistics*, **85** (1) 105–118.

Brueckner, J. K. and W. T. Whalen (2000) The Price Effects of International Airline Alliances, *The Journal of Law and Economics*, **43** (2) 503–546.

Bubalo, B. and J. R. Daduna (2011) Airport capacity and demand calculations by simulation—the case of Berlin-Brandenburg International Airport, *NETNOMICS: Economic Research and Electronic Networking*, **12** (3) 161–181, October 2011.

Bureau of Transportation Statistics (2016) National transportation statistcs, webpage, URL `http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/index.html`. Accessed on 20/04/2016.

Caceres, R., J. P. Wideberg and F. G. Benitez (2007) Deriving origin destination data from a mobile phone network, *Intelligent Transport Systems, IET*, **1** (1) 15–26, March 2007.

Calabrese, F., M. Diao, G. di Lorenzo, J. Ferreira and C. Ratti (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C: Emerging Technologies*, **26**, 301–313.

Cambridge Systematics Inc. (2013) Processing of California Household Travel Survey Data for Model Calibration and Validation, *Technical Report*, California High Speed Rail Authority, Los Angeles.

Chalasani, V. S. and K. W. Axhausen (2004) Mobidrive: a six weeks travel diary, *Travel Survey Metadata Series*, **2**, IVT, ETH Zurich, Zurich.

Chapin, F. S. (1974) *Human Activity Patterns in the City*, John Wiley & Sons, Hoboken.

Chen, C., L. Bian and J. Ma (2014) From traces to trajectories: How well can we guess activity locations from mobile phone traces?, *Transportation Research Part C: Emerging Technologies*, **46**, 326–337.

Chlond, B., J. Last, W. Manz and D. Zumkeller (2006) Long-Distance Travel in a Longitudinal Perspective: The INVERMO Approach in Germany, paper presented at the *85th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2006.

Chorus, C. G., T. A. Arentze and H. J. P. Timmermans (2008) A comparison of regret-minimization and utility-maximization in the context of travel mode choices, paper presented at the *87th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2008.

Cik, M., M. Fellendorf and J. Vogel (2014) Cell phone data as an extension for traffic models, *Straßenverkehrstechnik*, **11**, 739–744.

Crompton, J. L. (1992) Structure of vacation destination choice sets, *Annals of Tourism Research*, **19** (3) 420–434.

Crompton, J. L. and P. K. Ankomah (1993) Choice set propositions in destination decisions, *Annals of Tourism Research*, **20** (3) 461–476.

Crouch, G. I. (1994) The Study of International Tourism Demand: A Survey of Practice, *Journal of Travel Research*, **32** (4) 41–55.

Dagum, L. and R. Menon (1998) OpenMP: an industry standard API for shared-memory programming, *IEEE Computational Science and Engineering*, **5** (1) 46–55.

De Cantis, S., A. M. Parroco, M. Ferrante and F. Vaccina (2015) Unobserved tourism, *Annals of Tourism Research*, **50**, 1–18.

De Lapparent, M., K. W. Axhausen and A. Frei (2013) Long distance mode choice and distributions of values of travel time savings in three European countries, *European Transport*, **53**, 1–31.

de Palma, A., F. Marchal and Y. Nesterov (1997) METROPOLIS: Modular System for Dynamic Traffic Simulation, *Transportation Research Record*, **1607** (1) 178–184.

Department for Transport (2016) National travel survey: 2014, webpage, URL `http://www.gov.uk/government/statistics/national-travel-survey-2014`. Accessed on 17/02/2016.

Deville, J.-C. and C.-E. Särndal (1992) Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87** (418) 376–382.

Deville, J.-C., C.-E. Särndal and O. Sautory (1993) Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, **88** (423) 1013–1020.

Dillman, D. A. (2000) *Mail and Internet Surveys. The Tailored Design Method*, John Wiley & Sons, New York.

Draijer, G., N. Kalfs and J. Perdok (2000) Global Positioning System as data collection method for travel research, *Transportation Research Record*, **1719**, 147–153.

Dubernet, T. (2017) Explicitly correlating agent's daily plans in a multiagent transport simulation: Towards the consideration of social relationships, Ph.D. Thesis, ETH Zurich, Zurich.

Erath, A. L., A. Chakirov, P. J. Fourie, S. A. Ordóñez Medina, M. Shah, M. A. B. van Eggermond and K. W. Axhausen (2012) A large-scale agent-based transport travel demand model for Singapore: The implementation of MATSim, *Working Paper*, Future Cities Laboratory, Singapore-ETH Centre (SEC), Singapore.

Erhardt, G. D., J. Freedman, A. Stryker, H. Fujioka and R. Anderson (2007) Ohio long-distance travel model, *Transportation Research Record*, **2003**, 130–138.

ETH (2019) Scientific Compute Clusters, ETH Zurich, webpage, URL `http://https://www.ethz.ch/services/en/it-services/catalogue/server-cluster/hpc.html`.

Eurostat (2016) Tourism statistics - top destinations, webpage, URL `http://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics_-_top_destinations`. Accessed on 26/04/2016.

Fletcher, R. and M. J. D. Powell (1963) A rapidly convergent descent method for minimization, *The Computer Journal*, **6** (2) 163–168.

Flötteröd, G. (2017) A search acceleration method for optimization problems with transport simulation constraints, *Transportation Research*, **98**, 239–260.

Fourie, P. J. (2016) Synthesizing high-dimensional, agent-based transport demand data from two-dimensional aggregates with iterative multiple histogram matching, paper presented at the *15th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Singapore, May 2016. TRANSET (AAMAS 2016 workshop on TRANSportation applications of Equilibrium, incentives and game Theory).

Frei, A., T. Kuhnimhof and K. W. Axhausen (2010) Long-distance travel in europe today: Experiences with a new survey, paper presented at the *89th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2010.

Frick, M., B. C. Belart, M. Schmied, B. Grimm and D. Schmücker (2014) Zukunftsperspektiven der Langstreckenmobilität - Grundlagenstudie, *Final Report*, Institut für Mobilitätsforschung, Berne.

Frick, M. and B. Grimm (2014) Long-Distance Mobility: Current Trends and Future Perspectives, *Research Report*, Institut für Mobilitätsforschung, Munich.

Friedrich, M., K. Immisch, P. Jehlicka, T. Otterstätter and J. Schlaich (2010) Generating OD matrices from mobile phone trajectories, paper presented at the *89th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2010.

González, M. C., C. A. Hidalgo and A.-L. Barabási (2008) Understanding individual human mability patterns, *Nature*, **454**, 779–782.

Goulias, K. G., A. W. Davis and E. C. McBride (2018) Long Distance Travel and Destination Attractiveness, *Technical Report*, University of California Santa Barbara, Santa Barbara, June 2018.

Grimal, R. (2010) Mobilité à longue distance : plus de voyages s'effectuent en train, mais les seniors restent adeptes de la voitur, *Technical Report*, Setra, Paris, December 2010.

Hackney, J. K. and K. W. Axhausen (2006) An agent model of social network and travel behavior interdependence, paper presented at the *11th International Conference on Travel Behaviour Research (IATBR)*, Kyoto, August 2006.

Halat, H., A. Zockaie, H. S. Mahmassani, X. Xu and O. Verbas (2016) Dynamic network equilibrium for daily activity-trip chains of heterogeneous travelers: application to large-scale networks, *Transportation*, **43** (6) 1041–1059.

Hamed, M. M. and F. Mannering (1993) Modeling Travelers' Postwork Activity Involvement: Toward A New Methodology, *Transportation Science*, **27** (4) 381–394.

Hannes, E., D. Janssens and G. Wets (2008) Destination choice in daily activity travel: The mental map's repertoire, paper presented at the *87th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2008.

Harris, C. C., W. J. McLaughlin and J. D. Hunt (1994) Estimating total statewide travel, *Annals of Tourism Research*, **21** (4) 701–711.

He, M., M. He and S. Gao (2009) Destination choice modeling for home-based non-commute trips: Some improvements in utility function and a case study in China, paper presented at the *88th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2009.

Helbing, D. and S. Balietti (2012) How to Do Agent-Based Simulations in the Future: From Modeling Social Mechanisms to Emergent Phenomena and Interactive Systems Design, in D. Helbing (ed.) *Social Self-Organization - Agent-Based Simulations and Experiments to Study Emergent Social Behavior*, 135–159, Springer, Berlin.

Holz-Rau, C., J. Scheiner and K. Sicks (2014) Travel Distances in Daily Travel and Long-Distance Travel: What Role is Played by Urban Form?, *Environment and Planning A*, **46** (2) 488–507.

Hörl, S., M. Balac and K. W. Axhausen (2019) Pairing discrete mode choice models and agent-based transport simulation with MATSim, paper

presented at the *98th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2019.

Horni, A. (2013) Destination choice modeling of discretionary activities in transport microsimulations, Ph.D. Thesis, ETH Zurich, Zurich.

Horni, A., K. Nagel and K. W. Axhausen (2012) High-resolution destination choice in agent-based demand models, paper presented at the *91st Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2012.

Horni, A., K. Nagel and K. W. Axhausen (eds.) (2016) *The Multi-Agent Transport Simulation MATSim*, Ubiquity, London.

Hoteit, S., S. Secci, S. Sobolevsky, C. Ratti and G. Pujolle (2014) Estimating human trajectories and hotspots through mobile phone data, *Computer Networks*, **64**, 296–307.

Huntsinger, L. F. and R. Donnelly (2014) Reconciliation of Regional Travel Model and Passive Device Tracking Data, paper presented at the *93rd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2014.

IFSTTAR (2016) French national travel survey: 2008, webpage, URL `http://http://www.statistiques.developpement-durable.gouv.fr/transports/s/transport-voyageurs-deplacements.html`. Accessed on 17/02/2016.

Iovan, C., A.-M. Olteanu-Raimond, T. Couronne and Z. Smoreda (2013) Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies, in D. Vandenbroucke, B. Bucher and J. Crompvoets (eds.) *Geographic Information Science at the Heart of Europe*, 247–265, Springer, Cham.

Jakob, M. and Z. Moler (2013) Modular framework for simulation modelling of interaction-rich transport systems, paper presented at the *16th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, The Hague, October 2013.

Janzen, M. and K. W. Axhausen (2015a) Activity Planning in a Continuous Long-Term Travel Demand Microsimulation, paper presented at the *15th Swiss Transport Research Conference*, Ascona, April 2015.

Janzen, M. and K. W. Axhausen (2015b) Long-term-C-TAP simulation: Generating long distance travel demand for a full year, paper presented at the *94th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2015.

Janzen, M. and K. W. Axhausen (2017a) Destination and Mode Choice in an Agent-Based Simulation of Long-Distance Travel Demand, paper presented at the *17th Swiss Transport Research Conference*, Ascona, May 2017.

Janzen, M. and K. W. Axhausen (2017b) Holiday destination choice in a continuous activity-based microsimulation, paper presented at the *96th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2017.

Janzen, M. and K. W. Axhausen (2017c) Population Synthesis for Long-Distance Travel Demand Simulations, paper presented at the *6th Symposium of the European Association for Research in Transportation*, Haifa, September 2017.

Janzen, M. and K. W. Axhausen (2018) Simulating Continuous Long-Distance Travel Demand, paper presented at the *15th International Conference on Travel Behaviour Research (IATBR)*, Santa Barbara, July 2018.

Janzen, M., D. Charypar and K. W. Axhausen (2014) Long term simulation of a continuous target-based model, paper presented at the *14th Swiss Transport Research Conference*, Ascona, May 2014.

Janzen, M., M. Vanhoof, Z. Smoreda and K. W. Axhausen (2018) Closer to the total? Long-distance travel of French mobile phone users, *Travel Behaviour and Society*, **11**, 31–42.

Jin, X., H. Asgari and M. S. Hossan (2014) Understanding Trip Misreporting in Household Travel Surveys by Comparing GPS-Assisted and Diary-Based Samples, paper presented at the *14th COTA International Conference of Transportation Professionals*, Changsha, July 2014.

Johansson, M. V., T. Heldt and P. Johansson (2006) The effects of attitudes and personality traits on mode choice, *Transportation Research Part A*, **40** (6) 507–525.

Jonnalagadda, J., N. Freedman, W. A. Davidson and J. D. Hunt (2001) Development of microsimulation activity-based model for San Francisco: Destination and mode choice models, *Transportation Research Record*, **1777**, 25–35.

Karl, M., C. Reintinger and J. Schmude (2015) Reject or select: Mapping destination choice, *Annals of Tourism Research*, **54**, 48–64.

Klöckner, C. A. and E. Matthies (2004) How habits interfere with norm-directed behaviour: A normative decision-making model for travel mode choice, *Journal of Environmental Psychology*, **24** (3) 319–327.

Koppelman, F. S. (1989) Multidimensional Model System for Intercity Travel Choice Behavior, *Transportation Research Record*, **1241**, 1–8.

Koppelman, F. S. and V. Sethi (2005) Incorporating variance and covariance heterogeneity in the Generalized Nested Logit model: an application to

modeling long distance travel choice behavior, *Transportation Research Part B: Methodological*, **39** (9) 825–853.

Kuhnimhof, T., R. Frick, B. Grimm and P. Phleps (2014) Long distance mobility in central europe: Status quo and current trends, paper presented at the *European Transport Conference*, Frankfurt, October 2014.

Kuhnimhof, T. and C. Gringmuth (2009) Multiday multiagent model of travel behavior with activity scheduling, *Transportation Research Record*, **2134**, 178–185.

Kuhnimhof, T. and J. Last (2009) The Path to Better Long-distance Travel Data in Europe – The Potential of Combining Established Household Survey Instruments and Methodological Innovations Long Distance Mobility in Central Europe: Status Quo and Current Trends, paper presented at the *First International Conference on the Measurement and Economics Analysis of Regional Tourism*, San Sebastian, October 2009.

Lagarias, J. C., J. A. Reeds, M. H. Wright and P. E. Wright (1998) Convergence properties of the nelder-mead simplex method in low dimensions, *SIAM Journal of Optimization*, **9**, 112–147.

LaMondia, J. J., C. R. Bhat and D. A. Hensher (2008) An annual time use model for domestic vacation travel, *Journal of Choice Modelling*, **1** (1) 70–97.

LaMondia, J. J., T. Snell and C. R. Bhat (2010) Traveler Behavior and Values Analysis in the Context of Vacation Destination and Travel Mode Choices, *Transportation Research Record*, **2156**, 140–149.

Le Floc'h, F. (2012) Issues of Nelder-Mead Simplex Optimisation with Constraints, *SSRN Electronic Journal*, **1** (1).

Leeworthy, V. R., P. C. Wiley, D. B. English and W. Kriesel (2001) Correcting response bias in tourist spending surveys, *Annals of Tourism Research*, **28** (1) 83–97.

Lennon, J. J. (2003) *Tourism Statistics: International Perspectives and Current Issues*, Continuum.

Licoppe, C., D. Diminescu, Z. Smoreda and C. Ziemlicki (2008) Using mobile phone geolocalisation for 'socio-geographical'analysis of co-ordination, urban mobilities, and social integration patterns, *Tijdschrift voor economische en sociale geografie*, **99** (5) 584–601.

Limtanakool, N., M. Dijst and T. Schwanen (2006) On the participation in medium. and long-distance travel: A decomposition analysis for the uk and the netherlands, *Tijdschrift voor economische en sociale geografie*, **97** (4) 389–404.

Lu, L., Y. Xu, C. Antoniou and M. E. Ben-Akiva (2015) An enhanced SPSA algorithm for the calibration of Dynamic Traffic Assignment models, *Transportation Research Circular*, **51**, 149–166.

Madre, J.-L., K. W. Axhausen and W. Brög (2007) Immobility in travel diary surveys, *Transportation*, **34** (1) 107–128.

Märki, F. (2014) An Agent-Based Model for Continuous Activity Planning of Multi-Week Scenarios, Ph.D. Thesis, ETH Zurich, Zurich.

Märki, F., D. Charypar and K. W. Axhausen (2011) Continuous activity planning for a continuous traffic simulation, *Transportation Research Record*, **2230**, 29–37.

Märki, F., D. Charypar and K. W. Axhausen (2012a) Target driven activity planning, paper presented at the *91st Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2012.

Märki, F., D. Charypar and K. W. Axhausen (2012b) Validation of a continuous simulation model for daily travel, paper presented at the *12th Swiss Transport Research Conference*, Ascona, May 2012.

Märki, F., D. Charypar and K. W. Axhausen (2013) Integration of Household Interaction with a Continuous Simulation Model, paper presented at the *13th Swiss Transport Research Conference*, Ascona, April 2013.

Märki, F., D. Charypar and K. W. Axhausen (2014a) Agent-based model for continuous activity planning with an open planning horizon, *Transportation*, **41** (4) 905–922.

Märki, F., D. Charypar and K. W. Axhausen (2014b) Location choice for a continuous simulation of long periods under changing conditions, *Journal of Transport and Land Use*, **7** (2) 1–17.

Mellegard, E. (2011) Obtaining Origin/Destination-matrices from cellular network data, Master Thesis, Chalmers University, Gothenburg.

Miller, E. J. (1996) Microsimulation and activity-based forecasting, paper presented at the *Activity-Based Travel Forecasting Conference*, New Orleans, June 1996.

Miller, E. J. and M. J. Roorda (2003) A prototype model of 24-hour household activity scheduling for the Toronto area, *Transportation Research Record*, **1831**, 114–121.

Miller, E. J., M. J. Roorda and J. A. Carrasco (2005) A tour-based model of travel mode choice, *Transportation*, **32** (4) 399–422.

Moeckel, R., R. Fussell and R. Donnelly (2015) Mode choice modeling for long-distance travel, *Transportation Letters*, **7** (1) 35–46.

Montini, L., S. Prost, J. Schrammel, N. Rieser-Schüssler and K. W. Axhausen (2014) Comparison of travel diaries generated from smartphone

data and dedicated GPS devices, paper presented at the *10th International Conference on Transport Survey Methods*, Leura, November 2014.

Müller, K. (2017) A generalized approach to population synthesis, Ph.D. Thesis, ETH Zurich, Zurich.

Nagel, K. and G. Flötteröd (2009) Agent-based traffic assignment: Going from trips to behavioral travelers, paper presented at the *12th International Conference on Travel Behaviour Research (IATBR)*, Jaipur, December 2009.

Narayana, C. L. and R. J. Markin (1975) Consumer behavior and product performance: An alternative conceptualization, *Journal of Marketing*, **39** (4) 1–6.

Nelder, J. A. and R. Mead (1965) A simplex method for function minimization, *The Computer Journal*, **7** (4) 308–313.

Nelldal, B.-L. (1998) The experience of the SJ X2000 tilting train and its effect on the market, *Proceedings of the Institution of Mechanical Engineers, Part F*, **212**, 103–108.

Neumann, A. (2003) Korrekturverfahren für Stichproben von Verkehrsverhaltenserhebungen des Personenfernreiseverkehrs, Ph.D. Thesis, University of Natural Resources and Life Sciences, Vienna, Vienna.

Niininen, O., E. Szivas and M. Riley (2004) Destination loyalty and repeat behaviour: An application of optimum stimulation measurement, *International Journal of Tourism Research*, **6** (6) 439–447.

Nijland, L., T. A. Arentze and H. J. P. Timmermans (2014) Multi-day activity scheduling reactions to planned activities and future events in a dynamic model of activity-travel behavior, *Journal of Geographical Systems*, **16** (1) 71–87.

OECD (2012) OECD Tourism Trends and Policies 2012, webpage, URL http://www.oecd-ilibrary.org/industry-and-services/oecd-tourism-trends-and-policies-2012_tour-2012-en. Accessed on 26/04/2016.

Oom do Valle, P., J. A. Silva, J. Mendes and M. Guerreiro (2006) Tourist Satisfaction and Destination Loyalty intention: A Structural and Categorical Analysis, *International Journal of Business Science and Applied Management*, **1**, 25–44.

Oppermann, M. (2000) Tourism Destination Loyalty, *Journal of Travel Research*, **39** (1) 78–84.

Ordóñez Medina, S. A. (2015) Recognizing personalized flexible activity patterns, paper presented at the *14th International Conference on Travel Behaviour Research (IATBR)*, Windsor, July 2015.

Ordóñez Medina, S. A. (2017) Activity-based multi-agent simulation of urban transport for a week time horizon, Ph.D. Thesis, ETH Zurich, Singapore.

Ordóñez Medina, S. A. (2018) Inferring weekly primary activity patterns using public transport smart card data and a household travel survey, *Travel Behaviour and Society*, **12**, 93–101.

Outwater, M., M. A. Bradley, N. Ferdous, C. R. Bhat, R. M. Pendyala, S. Hess, A. J. Daly and J. J. LaMondia (2015a) Tour-Based National Model System to Forecast Long-Distance Passenger Travel in the United States, paper presented at the *94th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2015.

Outwater, M., M. A. Bradley, N. Ferdous, S. Trevino and H. Lin (2015b) Foundational knowledge to support a long-distance passenger travel

demand modeling framework, *Technical Report*, Transport Studies Unit, University of Oxford.

Pan, C., J. Lu, S. Di and B. Ran (2006) Cellular-Based Data-Extracting Method for Trip Distribution, paper presented at the *85th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2006.

Pendyala, R. M., R. Kitamura, A. Chen and E. I. Pas (1997) An activity-based microsimulation analysis of transport control measures, *Transport Policy*, **4** (3) 183–192.

Perdue, R. R. and M. R. Botkin (1988) Visitor survey versus conversion study, *Annals of Tourism Research*, **15** (1) 76–87.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (2007) *Numerical Recipes: The Art of Scientific Computing*, 3. edn., Cambridge University Press, Cambridge.

Pritchard, D. R. (2008) Synthesizing agents and relationships for land use / transportation modelling, Master Thesis, University of Toronto, Toronto.

Ranjan, G., H. Zang, Z.-L. Zhang and J. Bolot (2012) Are Call Detail Records Biased for Sampling Human Mobility?, *ACM SIGMOBILE Mobile Computing and Communications Review*, **16** (3) 33–44, December 2012.

Rich, J. and S. L. Mabit (2012) A long-distance travel demand model for Europe, *European Journal of Transport and Infrastructure Research*, **12** (1) 1–20.

Richardson, A. J., E. S. Ampt and A. H. Meyburg (eds.) (1995) *Survey Methods for Transport Planning*, Eucalyptus Press, Melbourne.

Rieser, M., D. Grether and K. Nagel (2009) Adding mode choice to multiagent transport simulation, *Transportation Research Record*, **2132**, 50–58.

Sarlas, G. and K. W. Axhausen (2015) Localized speed prediction with the use of spatial simultaneous autoregressive models, paper presented at the *94th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2015.

Sarlas, G. and K. W. Axhausen (2016) Exploring spatial methods for prediction of traffic volumes, paper presented at the *16th Swiss Transport Research Conference*, Ascona, April 2016.

Schaefer, A. and D. Victor (1997) The Past and Future of Global Mobility, *Scientific American*, **277** (4) 58–61.

Schlich, R. (2001) Analysing intrapersonal variability of travel behaviour using the Sequence Alignment Method, paper presented at the *European Transport Conference*, Cambridge.

Schonland, A. M. and P. W. Williams (1996) Using the Internet for Travel and Tourism Survey Research: Experiences from the Net Traveler Survey, *Journal of Travel Research*, **35** (2) 81–87.

Seddighi, H. R. and A. L. Theocharous (2002) A model of tourism destination choice: a theoretical and empirical analysis, *Tourism Management*, **23** (5) 475–487.

Smith, L., R. J. Beckman, D. Anson, K. Nagel and M. E. Williams (1995) TRANSIMS: TRansportation ANalysis and SIMulation System, paper presented at the *5th TRB National Transportation Planning Methods Applications Conference*, Seattle, April 1995.

Smith, R. S. and J. E. A. Wood (1977) Memory–Its Reliability in the Recall of Long Distance Business Travel, *Final Report*, Transport and Road Research Laboratory, Transport and Road Research Laboratory, Berkshire.

Smoreda, Z., A.-M. Olteanu-Raimond and T. Couronne (2013) Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment, in J. Zmud, M. E. H. Lee-Gosselin, M. A. Munizaga and J. A. Carrasco (eds.) *Transport Survey Methods: Best Practice for Decision Making*, chap. 41, 745–768, Emerald, Bingley.

Song, C., Z. Qu, N. Blumm and A.-L. Barabási (2010) Limits of predictability in human mobility, *Science*, **327** (5968) 1018–1021.

Southworth, F. (1981) Calibration of multinomial logit models of mode and destination choice, *Transportation Research Part A: Policy and Practice*, **15** (4) 315–325.

Spall, J. C. (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, *IEEE Transactions on Automatic Control*, **37** (3) 332–341, March 1992.

Swiderski, D. (1983) A model for simulating spatially and temporally coordinated activity sequences on the basis of mental maps, in S. M. Carpenter and P. M. Jones (eds.) *Recent Advances in Travel Demand Analysis*, 313–334, Gower, Aldershot.

Swiss Federal Statistical Office (BFS) and Swiss Federal Office for Spatial Development (ARE) (2012) *Mobilität in der Schweiz - Ergebnisse des Mikrozensus Mobilität und Verkehr 2010*, Swiss Federal Statistical Office (BFS) and Swiss Federal Office for Spatial Development (ARE), Neuchatel and Berne.

Swiss Railways (2016) The new Gotthard Tunnel – Switzerland through and through, *Final Report*, Swiss Railways, Berne.

Toole, J. L., S. Colak, B. Sturt, L. Alexander, A. Evsukoff and M. C. González (2015) The path most traveled: Travel demand estimation using big data resources, *Transportation Research Part C: Emerging Technologies*, **58, Part B**, 162–177.

Tourangeau, R. (1999) Remembering What Happened: Memory Errors and Survey Reports, in A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman and V. S. Cain (eds.) *The Science of Self-report: Implications for Research and Practice*, chap. 3, 29–47, Psychology Press, East Sussex.

Ullman, H. and L. Aultman-Hall (2018) Exploring Vermonter's Self-Reported Unmet Long-Distance Travel Need, paper presented at the *97th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2018.

Van Nostrand, C., V. Sivaraman and A. R. Pinjari (2013) Analysis of long-distance vacation travel demand in the United States: a multiple discrete-continuous choice framework, *Transportation*, **40** (1) 151–171.

Vovsha, P. (1997) Cross-Nested Logit model: An application to mode choice in the Tel-Aviv metropolitan area, *Transportation Research Record*, **1607**, 6–15.

Wang, P., T. Hunter, A. M. Bayen, K. Schlechtner and M. C. González (2012) Understanding Road Usage Patterns in Urban Areas, *Scientific Reports*, **2** (1001) 1–6.

Weckström-Eno, K. (1999) Long distance passenger travel, *Final Report*, European Commission and Eurostat, Luxembourg.

White, J. and I. Wells (2002) Extracting origin destination information from mobile phone data, paper presented at the *Eleventh International Conference on Road Transport Information and Control, 2002. (Conf. Publ. No. 486)*, London.

Wolf, J., M. Oliveira and M. Thompson (2003) The impact of trip underreporting on vmt and travel time estimates: Preliminary findings from the california statewide household travel survey gps study, paper presented at the *82nd Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2003.

Yim, Y. (2003) The State of Cellular Probes, *Final Report*, **UCB-ITS-PRR-2003-25**, University of California, Berkeley, Berkeley, July 2003.

Zheng, J. and J. Y. Guo (2008) Destination choice model incorporating choice set formation, paper presented at the *87th Annual Meeting of the Transportation Research Board*, Washington, D.C., January 2008.

Zimmermann, A., K. W. Axhausen, J. Beckmann, K. J. Beckmann, M. Düsterwald, E. Fraschini, T. Haupt, A. König, A. Kübel, G. Rindsfüser, R. Schlich, S. Schönfelder, A. Simma and T. Wehmeier (2001) Mobidrive: Dynamik und Routinen im Verkehrsverhalten: Pilotstude Rhythmik, *Final Report*, Federal Ministry of Education and Research, PTV and IVT, ETH Zurich and Institut für Stadtbauwesen und Strassenverkehr (ISB), RWTH Aachen, Karlsruhe, Zurich and Aachen.

Zmud, J., M. E. H. Lee-Gosselin, M. A. Munizaga and J. A. Carrasco (eds.) (2013) *Transport Survey Methods: Best Practice for Decision Making*, Emerald, Bingley.

Zumkeller, D., W. Manz, J. Last and B. Chlond (2005) Die intermodale Vernetzung von Personenverkehrsmitteln unter Berücksichtigung der

Nutzerbedürfnisse (INVERMO), *Final Report*, Institut für Verkehrswesen, University of Karlsruhe, Federal Ministry of Education and Research, Karlsruhe.