PAUL BOUMAN

Passengers, Crowding and Complexity

Models for Passenger Oriented Public Transport

Passengers, Crowding and Complexity

Models for passenger oriented public transport

Passengers, Crowding and Complexity
Models for passenger oriented public transport

Passagiers, druktevorming en complexiteit
Modellen voor passagiersgeoriënteerd openbaar vervoer

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board

The public defense shall be held on

Thursday 15th of June 2017 at 15:30

by

PAUL CORNELIS BOUMAN
born in De Bilt, the Netherlands.

Erasmus University Rotterdam

Doctoral Committee

Promotors: Prof.dr. L.G. Kroon († 14-09-2016)
Prof.dr. A. Schöbel
Prof.dr.mr. P.H.M. Vervest

Other members: Dr. J. van Dalen
Prof.dr. D. Huisman
Prof.dr. K. Nagel

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: `http://repub.eur.nl/`

ERIM PhD Series in Research in Management, 420
ERIM reference number: EPS-2017-420-LIS
ISBN 978-90-5892-482-7

Design: PanArt `www.panart.nl`

Cover image design: Emily Simpson

This publication (cover and interior) is printed by Tuijtel on recycled paper, BalanceSilk®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Ecolabel, FSC®, ISO14001.
More info: `www.tuijtel.com`

Acknowledgements

It can be argued that the origin of this thesis can be traced back to a late Thursday evening, somewhere in the spring of the year two thousand and eleven at the bar of Biton, a humble student association in the city of Utrecht, The Netherlands. During this fateful evening fellow student Marijke Bodlaender informed me about a vacancy for a PhD Student at the Erasmus University, which she received via the mailing-list on Discrete Mathematics and Algorithms. This information led to an e-mail to and a phone call from Leo Kroon, a very pleasant job interview with Peter Vervest and Evelien van der Hurk and ultimately, the thesis you are currently reading.

Over the course of the last couple of years I always believed that writing these acknowledgments would be a nice way to end the writing of my thesis, yet I did not realize that the amount of space available is not enough to express my full gratitude. Words cannot do justice to the many stories, memories and experiences. On many occasions the cliché “You had to be there” is amazingly appropriate.

Academic Acknowledgments First, I want to state I am very grateful for the support of Leo Kroon during the years I was working on my thesis. He was always patient to explain the intricacies of railway operations and suggest which questions are of interest to Netherlands Railways. He had an uncanny ability to find unclarity and spelling mistakes in my writing. As a person, I appreciated his sense of humor and the fact that I could often walk into his office unannounced to give him an update on what I had been working on. It saddens me that I had to finish my thesis without his guidance and I hope he found the peace he desired.

I am also grateful to Peter Vervest. I did not expect to find his sincere enthusiasm for the exact sciences in a business school, but it has lead to more pleasant conversations than I can count. I also appreciate his ability to challenge people to think outside the box, even though the visions of the future he paints to achieve this are sometimes a bit unorthodox. His effort in pushing me to finalize my thesis has been indispensable.

Anita Schöbel was kind enough to step in the role as second promotor after Leo's passing. I appreciate her helpful feedback and really enjoyed working with her and Marie Schmidt on the chapters in Part III. Now that my thesis is finished I hope we can continue working on these models. Jan van Dalen facilitated the opportunity to execute the experiment from Chapter 3 during his course and provided valuable guidance during its execution. I look back fondly at the joint lunch conversations while working at RSM. His suggestions related to typography and style helped to make this a beautiful thesis. I also thank Kai Nagel for his advice regarding simulations and complexity science. From the early stages of the project I have been impressed with MATSim: The open nature of the project together with the very modular setup involves a quality in software engineering I have seldom seen in a scientific context. The final member of the inner committee, Dennis Huisman, had numerous helpful comments that improved this thesis a lot. I thank him for his efforts.

As the "Complexity in Public Transport" project provided funding for two PhD students, I am happy that Evelien van der Hurk was the other PhD student involved in the project. I could always discuss ideas with her that pushed my research forward. I am grateful for the friendship that developed over the years and am happy that she accepted to be my paranimf. Furthermore, I am happy that Clint Pennings accepted to be my paranimf. I am fond of his enthusiasm for his latest discoveries and hobbies, and his ability to make a whole afternoon fly by with (nerdy) conversations.

I am also thankful for the many carpool sessions with Luuk that messed up the smart card data of NS by making me randomly appear at Woerden station. Evşen was a fabulous office mate and sushi queen. I hope to watch many more art-house movies with her, with or without an "explicateur". I am happy Nick was willing to take a picture of me for this thesis and I enjoyed sharing an office with him for a while. I also shared an office with Wouter for some time and I enjoy his enthusiasm for all kinds of games and complexity science. I thank Joris for the work on the RAS competition and traveling to Tokyo and Nagano with Luuk and me.

I am also happy to have worked and continue working with Marie on many interesting and challenging optimization problems. Her mathematical rigor has prevented me to jump to conclusions too soon on a number of occasions. I deeply respect the way Niels is able to bridge managerial practice with algorithmic research and am fond of his sense of humor. I really enjoy working on the drone related research with both of them and it still feels like an escalated "hobby"-project.

I also thank my other former office mates, Irina and Gábor, for being such pleasant company. I have high praise for Sarita's role as the "mother" of the Departyment. Furthermore, I thank Bas, Christina, Eric, Erwin, Gert-Jaap, Jelmer, Jørgen, Konstantina, Lars, Martin, Milan, Ni-Luh, Micha, Panos, Paul, René, Rob, Roelof, Thomas, Tim, Ting and many others for the nice conversations and events. Whenever I needed to arrange something, Cheryl and Ingrid were able to help out and their contribution to the enjoyable atmosphere at the TOM department is paramount.

When Rommert Dekker and Patrick Groenen invited me to teach programming classes at the Econometric Institute, I did not imagine it would lead to a tenure track position. I am thankful for their trust and support. The helpful suggestions and tips of Erik, Flavius, Paul and Wilco made teaching my first course a smooth affair. Having Gert-Jan as a co-teacher was also fantastic: We can be very proud of what we achieved and I am sure we established a sturdy foundation for other courses.

I was able to make a steady transition to the Econometric Institute, as I knew many of its members from seminars, conferences and other academic events. Judith and Twan, thank you for always being up for discussing mathematical puzzles on a whiteboard, even if you have no time to do so. I also thank the other (former) members of the Econometric Institute for the enjoyable working environment. Albert, Charlie, Chiel, Evelot, Harwin, Kevin, Kim, Kristiaan, Marieke, Mathijn, Nemanja, Remy, Richard, Rowan, Sha, Shadi, Thomas Visser, Thomas Breugem, Willem, Zahra and others: I am sure to keep working with many of you and I promise to write more personal acknowledgments on the next occasion. I thank Anneke, Lidewij, Marianne, Marjon and Renee for always helping me out with administrative issues in a satisfying way.

I was also welcomed at Netherlands Railways, where conversations with Erwin, Hilbert, Joël, Pieter-Jan, Ramon and the others were always a nice change of pace. Furthermore, there were many people I met during and around conferences. Although I do not list them by name, I look back fondly at the many adventures and conversations and hope to see many of them at future events. To finish, I thank Han and Marjan for introducing me to academic writing and introducing me to the fascinating world of Catan expansions and the Eurovision Song Contest.

Formal Acknowledgments First and foremost, I thank the Netherlands Organization for Scientific Research (NWO) for funding the "Complexity in Public Transport" project, grant #600.645.000.08, which paid for my research. I also thank the Erasmus Research Institute of Management for their support in allowing me to present my research at various conferences. I am thankful to the anonymous reviewers of Chapters 2, 4, 5, 7 and 8 for their valuable feedback. Additional thanks go to Kai Nagel and his colleagues for their comments and support with the simulations in Chapter 4. The work on Chapter 7 was additionally supported by grant SCHO 1140/3-2 within the DFG programme *Algorithm Engineering*.

Personal Acknowledgments Finally, there are many friends outside of academia to whom I am grateful. I hope to spend more time with many them now that this thesis is finished. Here is a humble attempt to list the people I wish to thank.

Meral gave me an early introduction to the spirit of Rotterdam and I highly value her friendship. Rianne is one of the most relaxed, and secretly also geeky, persons I know. I appreciate Maal's ongoing artistic endeavors and I look forward to attend more of her performances. When I go shopping for groceries on a Sunday, Trees' cheerful demeanor reminds me of many great parties. We all miss Evelien and I should remember her enthusiasm and somewhat peculiar sense of humor more often.

I enjoy the company of my pub quiz team. Alex is, besides a fantastic cook and fellow film enthusiast, the steadfast captain of the team. I thank Wilmar for being a connoisseur of the ridiculous and putting these qualities to good use by inventing things like the “mini kapsalon”. Margot’s knowledge of music and Chemistry proves useful on many occasions. When it comes to history questions, Machil is often up for the job and his knowledge of pretty bad contemporary music always catches me by surprise. Finally, it is clear Kees is made for his job as he knows all news facts and gossip, and knows how to tell a couple of excellent stories during the breaks.

Thanks to Janneke I am reminded every once in a while to go get a haircut. On top of that she was a fantastic host in Barcelona. I now know the difference between good and bad Sangria de Cava. I thank Christie for dragging me along on this trip.

There is also the illustrious society of port appreciating young ladies and associates. If it was not for Nora I would have been killed by virtual zombies many times. Willemijn’s skill to photograph Lego figurines in funny poses put a smile on my face on numerous occasions. Willemien often has amazing stories of tiny adventures and discoveries which I thoroughly enjoy. I miss having discussions about Mathematics on beer coasters with Koen and I fear I will not be able to keep up when he is done with his PhD in Mathematics. Thanks to Dieuwertje I was able to discuss PhD life with someone from a totally different university in a different country. Thanks to her research I realize how difficult it can be to compute the actual costs often assumed as given in many optimization models. I also thank Bas, Dexx, Manon and Maria for many enjoyable conversations.

I am happy to have chosen Emily for the design of the cover of this thesis: I love it. The typical sense of humor and RPG Maker games of Casper remain an inspiration to me. I envy Julian’s relentless skill with Pink Gold Peach in Mario Kart 8. I thank Bas and Friso for the many adventures in the Borderlands and beyond. I thank Manon for the intellectual discussions about privacy and copyright, Sophie for the enjoyable conversations about education in the super market at closing time and Elisa for her endless enthusiasm, kindness and garbage disposal advice. Together with Alex and Janneke, whose sharp sense of humor is a blast, I hope for many more comedy nights. Finally, I thank Lennert for never playing the drums too loud and his world-class skill of rolling down a hill, Bas who keeps on rocking `\m/` and Wijb who might one day invent how to power an accordion with thermal energy. I also hope there will be more opportunities for us to play board games or see a movie with Eva, Liesbeth, Steven and Tessa in the near future, when the dust of having babies settles.

Last, but surely not least, I thank my family. Willy, Liset, Linda, Oma, Jannie, Rein, Herman, Henriëtte and Ralph: Thanks for all the support during these years.

Utrecht, 15th of May 2017
Paul Bouman

Contents

1 Introduction

We see a car briefly driving forward in a tunnel before it stops, completely surrounded by other cars. As the camera pans to the other cars, we see people waiting patiently for the situation to change. While the car slowly fills with smoke, we see a bus so full with passengers that they hang their arms out of the windows, while all traffic remains motionless. The driver of the car climbs out of a window and soars away out of the tunnel.

This visualization of congestion and crowding to convey an uncomfortable, nightmarish feeling in the famous opening sequence of *Otto e Mezzo* (Fellini, 1963), was so effective that it has been imitated numerous times in later movies (Allen, 1980; Schumacher, 1993). In the same year Fellini established the cinematic language of peak hour congestion, future Nobel laureate William S. Vickrey, who is by some considered to be the father of *congestion pricing*, published a paper "*Pricing in Urban and Suburban Transport*" Vickrey (1963) that argues that the absence of peak-hour pricing in urban transportation is irrational and wasteful. Vickrey explains that nearly all costs of transportation systems are determined at the peak level and suggests that pricing strategies should be employed to promote economic use of transport systems rather than directly financing the services. The paper also proposes a technical system to realize differentiation pricing, which turns out to be very similar to the smart card ticketing systems employed by many operators in the late 1990's and the early 2000's. In the same decades, multiple proposals related to peak hour pricing of car traffic have been discussed in the parliament of the Netherlands, but none of them have made it into laws.

In 2015, crowded trains become an increasing nightmare for the board of Netherlands Railways (NS), as old rolling stock is taken out of circulation before new rolling stock is available (NOS, 2015). Furthermore, governmental benefits for students traveling by public transport are increased while a decrease was expected. In an attempt to control the damage, the CEO of NS proposes that universities change their schedule to move students out of the peak hours (as students are a significant flow of passengers in the Netherlands). As a result some Dutch student unions (Oosterhuis and van Meteren, 2016) protested, arguing that that NS makes the problems of the company the problems of students and universities without any consultation with them. After a television interview about the problems, the CEO of NS is ridiculed on television in a popular Dutch satirical news program (Lubach, 2015).

In the Netherlands, the pricing of peak demand in both public and private transportation has been a topic of heated debate for decades, without a strong consensus in the Netherlands. NS has implemented an off-peak discount subscription service, the "*Voordeelurenkaart*", that provides travelers a forty percent discount outside of the peak hours. This shows that peak hour demand is very costly for NS. As rolling stock and infrastructure makes up a large part of the costs of public transport, it can be easily reasoned that the peak demand determines the majority of the costs. Suppose a hundred new passengers show up in an existing public transport system, it is easy to argue that it is cheaper to let them travel with train services that have spare capacity than to invest many millions in a new unit of rolling stock – especially as one could argue that it is unfair to existing passengers to let them pay for this investment. However, as passengers do not always have the flexibility to travel at the moment at which the fares are the cheapest, some might argue that peak hour pricing hurts hard working citizens the most, while the unemployed benefit a lot. As a result, the Dutch government imposes a limit on the amount of money a single commuter may spend yearly on the costs of travel. It is thus not difficult to see that experiments with pricing incentives during the peak hours are a delicate subject from a political and public relations perspective.

Meanwhile, travelers and public transport operators have to cope with crowded vehicles and traffic congestion on a daily basis. With the increased availability of information technology, more and more individuals have some flexibility to do work at home or in the train, use video conferencing with persons at different locations and strategically plan their journeys at times when crowding or congestion is not too severe. At the same time, public transport operators now can use smart card data to observe the travel decisions of their passengers in greater detail than ever before, while they can use advanced decision support tools and information technology to influence the operational processes faster than before. This provides new opportunities for load balancing beyond mere congestion pricing or revenue management policies, as providers can attempt to influence passengers with suggestions about under utilized services and modify rolling stock capacities adaptively based on real time data.

In this thesis we apply ideas from *complexity science* to study how we can make an effort to improve public transport systems by using the opportunities created by information technology with relation to crowding and capacity. The main research goal of this thesis is formulated as follows:

Research Question. *Can micro-data on individual travel behavior be used to improve capacity utilization in daily public transport operations - why and how?*

The new opportunities introduced by information systems will result in more and faster interactions between travelers and the public transport operators. Such interactions are typically hard to take into account during strategic planning and scheduling, as many of the traditional economic models assume a transport system will converge to an equilibrium, while many of the models for line planning and rolling stock scheduling model demand from passengers using static high level aggregated flows. The modeling approaches typically associated with *complexity science* are useful to study these out-of-equilibrium dynamics and model day-to-day decision behavior based on observations made through smart card ticketing systems. The approaches take into account that public transport operators get more and more data related from smart card ticketing systems and other sources introduced by an increasing adoption of information systems.

Readers Guide This chapter gives the reader an overview of the structure of this thesis. The remainder of this chapter is organized as follows: In Section 1.2 we introduce complexity science, in Section 1.1 we introduce the "Complexity in Public Transport Project" that includes this thesis, in Section 1.3 we introduce the research challenge and goals of this thesis, in Section 1.4 we introduce and summarize the chapters and results of this thesis, in Section 1.5 we discuss the authorship of the chapters in this thesis and in Section 1.6 we conclude and provide some directions for future research.

1.1 Complexity in Public Transport

The work on this thesis was executed within the "Complexity in Public Transport" (ComPuTr) project (Kroon and Vervest, 2010) that was conducted within the Complexity program of the Netherlands Organization for Scientific Research (NWO). The Complexity program was a multi-disciplinary program that ranges from arts and humanities to physics and from economics to life sciences and called for applications of complexity science in different areas of society. An overview of the results of all projects in the program and directions for future research are presented in the NWO booklet "Grip on Complexity" (NWO, 2014).

Three common characteristics prevalent in all complex systems, including these studied in the different research projects of the NWO Complexity program, are defined by Boccara (2010) as 1) it consists of a large number of interacting agents, 2) it exhibits emergence, i.e. self-organizing collective behavior and 3) there is no central controller that dictates this behavior. Within public transport, the travel decisions of passengers are not controlled by public transport operators, although the operations often are. Therefore an important goal of the ComPuTr project was to include passengers and their behavior in planning and dispatching models. In Kroon and Vervest (2010), the main research goal of the ComPuTr project was formulated as follows:

> How to analyze and model (massive data on individual) passenger behavior in complex public transport networks? As a result: To define how public transport operators can apply this most effectively, both in the planning stage and in real-time operations?

Within the ComPuTr project two PhD research projects were formulated. The first project focuses on real-time operations, specifically on disruption management. Evelien van der Hurk successfully completed her PhD thesis on this topic in the Summer of 2015 (van der Hurk, 2015). The second project is completed in this thesis and focuses on the planning stage with a focus on peak demands and capacity.

As a first step to study peak demand, we converted smart card data recorded in a public transport system into an agent population for the large scale agent-based simulation framework MATSim (Horni et al., 2016). MATSim models agents based on the paradigm of activity based modeling (Axhausen and Gärling, 1992; Miller, 1997). While the traditional four-step model (McNally, 2000) for demand forecasting uses macroscopic properties such as land use and demographics to estimate travel volumes of travel demand between pairs of regions, activity based modeling aims to model individual households or even people. The possibility to model travelers at a microscopic level opens up the opportunity to model individual behavior and complex interactions between individuals. At the same time, the macroscopic data used by the four-step model can be used to synthesize a population. The disadvantage of activity based modeling is that the increased model details results in greater requirements for computational power. However, with modern computers this is not a very serious concern, as can be witnessed from the many large scale simulation scenario's discussed in the MATSim book (Horni et al., 2016).

Based on the results of this first step, we concluded that the main challenge in studying crowding interactions with a very realistic simulation model such as MATSim requires careful modeling of economic behavior of passengers. As the research team preferred to leverage expertise within the realm of complexity science, it was decided to focus on simple models that capture the essence of crowding in favor of developing a very precise model of economic behavior.

During one of the Winter Schools organized by NWO for participants in the Complexity program, we learned about the principle of the El-Farol bar game (Arthur, 1994). This model has a couple of striking features: It deals with crowding and it assumes individuals decide individually and independently whether they will occupy the capacity of the system, while the economic behavior was selected at random from a set of many predefined strategies. Although this model had been applied in the context of route choice for car traffic (Bazzan et al., 2000; Wahle et al., 2000), the application of the model in the context of time-choice and dynamic capacity optimization in public transport was not explored at the time. The first part of this thesis focuses on advances due to extensions of this model to the context of public transport.

The second part of the thesis focuses on the analysis of passenger behavior, based on observations collected via smart card ticketing systems. We develop algorithms that can detect activity patterns that are relevant in the context of capacity management and time choice modeling. This includes activities such as being at home or at work, but also shopping. We provide new models to find such patterns and introduce a way to validate the quality of the methods that produce such patterns based on smart card data.

The final part of the thesis focuses on the decentral aspect of passenger choice. In many situations, passengers may not have complete information on the state of public transport; either because the operator does not provide the information or because the operator does not have the information themselves (this is often the case when disruptions occur). We analyze how individual passengers can still make decisions if information is absent and identify different goals the passengers may have in mind while making such decisions. For example, a passenger who travels home after a day of work wants to arrive as soon possible, but a passenger who travels to a music show or to the airport to catch a flight wants to arrive before the show starts or flight departs. The proposed decision strategies can be applied for these different objectives.

1.2 Complexity and Complex Systems

In the explanation of the ComPuTr project, we introduced three common characteristics of complex systems. The science of complex systems has become very popular since the first research institute focused solely on complexity science, the Santa Fe institute, was established in 1984. Years later this popularity was one of the catalysts that lead to the unprecedented multidisciplinary call for proposals by Netherlands Organization for Scientific Research (NWO) under which the ComPuTr project was accepted. The common characteristics that occur in complex systems are that a system has *many interacting parts*, which are sometimes also called *agents*, that the behavior of the population of these parts is surprising and difficult to predict based on the behaviors of the individuals and that the behavior of the system is not dictated from a centralized controller.

Various modeling approaches are often associated with complexity science. One recommended introductory textbook on the modeling of complex systems (Boccara, 2010) covers *differential equations*, *maps*, *cellular automata*, *agent-based models* and *complex networks*. From these models, researchers have observed a wide array of phenomena such as *emergent behavior*, *power-laws*, *bifurcations* and *self-organized criticality*. These concepts are especially useful to compare phenomena and find commonalities in various disciplines and applications and as a result these concepts serve as a *common* *language*, that allows researchers from different backgrounds and disciplines to work across the traditional boundaries between scientific disciplines.

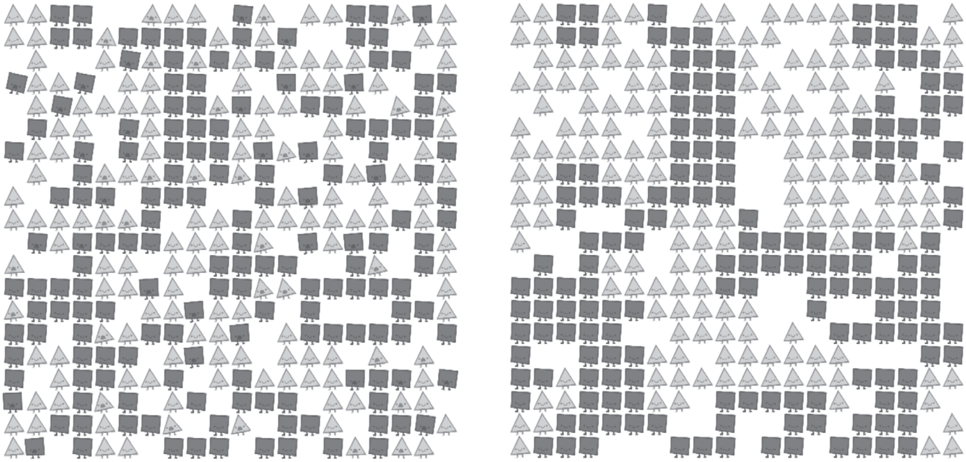
With this long list of modeling approaches and phenomena, it comes at no surprise that it is very difficult to give a clear and succinct definition of complexity science. While, for example, *physics* can be described as the "*study of nature*", *chemistry* as the "*study of matter*" and *biology* as the "*study of life*", a similar description for complexity science could be the "*study of interaction and emergent phenomena*", which is not the most helpful description for the average layperson.

An important side of complexity science is its relationship with other disciplines. Here at least two important aspects come to mind. First, while interaction is at the core of complexity science it is also important in many other disciplines and subjects. One famous example, the Lotka-Volterra model for predator-prey interactions (Lotka, 1910) was originally proposed in the context of ecology, but it is also suitable to study interactions in the economy (Goodwin, 1982). While this shows the strength of complexity science to study such commonalities across scientific disciplines and investigates the general principles underpinning them, it must be acknowledged that many of the models employed were developed by researchers who are often not active within the complexity science community. It is thus useful to be aware that in many cases complexity science is a discipline that connects other disciplines.

Secondly, the traditional *scientific method*, where we formulate a model, make a prediction based on the model and then validate the model (or reject it) based on the accuracy of how the prediction compares to empirical observations, assumes that models are always capable of making good predictions. Complexity science often deals with models that abstract away a lot of details and realism of the actual systems and focus on how aggregate behavior that emerges from the interactions in the system. Such models are not always useful for making predictions of real life systems and as a result it can be very difficult to perform validation. The advantage is that complexity science is able to address questions and issues that cannot be approached with purely analytical methods, but can be simulated on a computer. For a brief overview of some of the differences between complexity science and traditional science and engineering, we refer the reader to the introduction of Downey (2012).

There are numerous examples of interesting models and every time an author introduces complexity science they present their personal favorites. In this case we will start with the Schelling segregation model (Schelling, 1971), which has agents of two types (e.g. triangular and square) living on a cellular grid. An agent is happy if at least 30% of their neigbours on the grid is like them. Every round, an unhappy agent moves to a random unoccupied cell in the grid world. This is repeated until no unhappy agents exist. Although 30% seems like a requirement that is easily met, it still leads to severe segregation on the level of the population, as is displayed in Figure 1.1. An interactive "playable post" by Vi Hart and Nicky Case which does an outstanding job in explaining the model is suggested in the caption of Figure 1.1.

The main message of the model is that even though we have individual decision rules that seem reasonable this may result in undesirable emergent behavior at the



(a) Random distribution of shapes (b) The same population after convergence

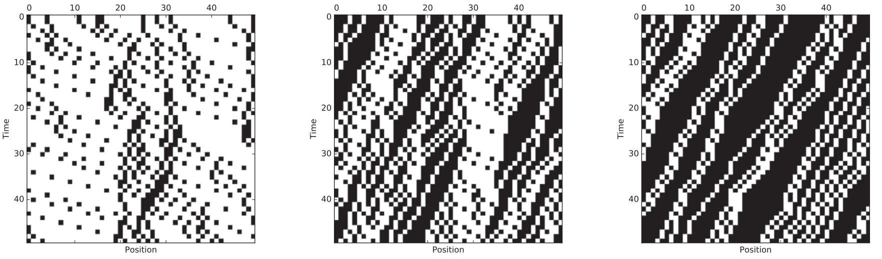
Figure 1.1: The Schelling Segregation model as presented in the "Parable of the Polygons" by Vi Hart and Nicky Case (https://ncase.me/polygons). Shapes are unhappy if less than $\frac{1}{3}$ of their neighbors have the same shape and move to a random empty spot in case they are unhappy.

population level. The segregation at the population level is more severe that can solely be explained as the "*sum of the parts*"; the interactions between multiple agents play a pivotal role here.

The second example of complexity science is the Nagel-Schreckenberg model (Nagel and Schreckenberg, 1992). In this model, cars drive in a sequence of cells from left to right. Each car has a speed and tries to accelerate to a maximum speed (4 cells per step in our example), increasing their speed by one step each round. In case a car would hit the car in front of them, they stop immediately to avoid a collision. With a small probability (0.2 in our example) a car randomly breaks and decreases its speed by one. The initial distribution of cars along the road is a uniform shuffle according to a predefined density. If we vary the density, we can observe the impact the density and the above rules have on the flow of the traffic, as can bee seen in Figure 1.2.

If the density of cars is low (as it is in Figure 1.2a), we can see that over time cars move to the right. This is indicated by the dotted lines moving to the bottom right of the figure. In case of a very congested road (as in Figure 1.2c) we observe traffic jams that slowly move backward instead of forward! Even though the cars themselves go right (albeit in a very slow manner), the emerging behavior of the traffic jams is surprising, and cannot be simply explained from the behavior of an individual car.

Sometimes models such as these are criticized for being overly simple. People come in more flavors than two and typically do not move to a random location.



(a) 20% car density (b) 50% car density (c) 70% car density

Figure 1.2: Three simulations of the Nagel-Schreckenberg model with varying car densities. Cells on the road containing a car are displayed as black cells, while empty cells of the road are white. Every line of the figures is the state of a road at a certain step in time. Time moves from top to bottom. We can see that if the cars flow freely (on the left) they move to the right. In case of traffic jams (to the right), the traffic jams themselves move backward.

Cars do not drive on a grid-based road and highways usually have more than one lane. However, the strength of these models is precisely that they are very simple, but still have surprising or counter-intuitive emergent properties. Understanding these models helps us to guide empirical research by suggesting what to look for: Do traffic jams occur out of nowhere on real highways in a similar way as the Nagel-Schreckenberg model describes? Can social phenomena such as segregation or lack of diversity be explained by a minor preference to be among similar people instead of big conspiracies against minorities?

In this thesis, we aim to use complexity science in a similar way. By using a simple, agent-based model and simulation, we gain insights that help us formulate new questions that help us get a grip on crowding behavior in public transport systems.

1.3 Research Goals

In order to address the main research question of this thesis, we define three steps that we will address separately. The focus of each of these steps is as follows:

1. **Modeling** of crowding caused by interactions between passengers and by interactions between the passengers and the public transport operator.

2. **Data-driven analysis** of travel patterns. In order to build more realistic models that incorporate crowding behavior, we must transform smart card data from raw journeys to activity types that are useful for crowding analysis.

3. **Decentralized** decision models that allow us to implement agents that can make travel decisions even if there is uncertainty about capacities or due to disruptions.

We will consider each step in more detail by formulating a research goal that corresponds to one of the steps and a part of this thesis. The research goal for the first step is formulated as follows:

> **Thesis Research Goal I.** *Propose and analyze a simple model that can describe emergent crowding behavior occurring due to the interactions of passengers and the public transport operator.*

The main strength of *simple* models is that they allow us to investigate the dynamics of systems. Models such as the Nagel–Schreckenberg model, or the Schelling Segregation Model give us insight in the behavior of a system over time. When we work with simulation, there is always a trade off to make: If we have only a few rules we can be certain that the observed behavior must be caused by these rules, but the resulting model is often too simple to be immediately applied in practical situations. If we create a very realistic model with many rules and parameters, the model is more suitable to support decisions and make predictions, but it is very difficult to attribute observed behaviors to specific rules used in the simulation. As our knowledge about the interaction between a public transport operator and passengers is fairly limited, it is from a scientific point of view sensible to focus on simple models.

Inspired by the paper that introduced the El-Farol Bar Game (Arthur, 1994) and the related work on minority games (Challet et al., 2004), we had a starting point to develop an extension that can capture crowding interactions. While similar models have been applied in transportation before (Bazzan et al., 2000; Wahle et al., 2000, 2002), this has always been within the context of car traffic, where crowding occurs between individual vehicles on a road where departure times can be chosen freely. The crowding dynamic in public transport is different as it occurs within the same vehicle and mostly has impact on comfort, while departure times are dictated by a schedule.

The second research goal of the thesis focuses on smart card data and investigates how we can apply it beyond the traditional approaches that estimate only macroscopic passenger flows.

> **Thesis Research Goal II.** *Develop algorithms that can extract temporal and spatial patterns relevant to agent-based modeling from journey sequences such as smart card data.*

Current research related to smart card data focuses mostly on estimating origin-destination flows, but not on behavior of individual (groups of) passengers. A 2011 survey paper on smart card data mentions the "*total disaggregate approach*" as a viable area of research (Pelletier et al., 2011). Using data collected by the public transport systems, typically via smart card data or smart phone applications, we can try to gain insight into aspects of passenger behavior that are relevant to the models we consider. Although operators have options to measure or estimate the volume of passengers that move between two locations on an average day (Trépanier et al., 2007), and numerous models can predict which travel option passengers will pick when confronted with multiple alternatives (Ben-Akiva and Lerman, 1985), decision behavior over time is not as well understood. The longitudinal nature of the data collected with modern devices allows us to analyze consecutive choices and include insights from these studies in our simulation.

One pitfall that we stumble upon from time to time is to think along the following line: *with smart card-data, we know everything about the passenger there is to know*. Unfortunately this is wishful thinking, as the data only tells us what the passenger decided in the situation that occurred in the past, but does not tell us at all what the passenger would have decided if that situation was different. In the transportation research community, this ties in with a relatively old discussion of *stated choice* versus *revealed choice* (Wardman, 1988; Ben-Akiva et al., 1994)

The final research goal of the thesis addresses the question how we can model decisions of decentralized passengers that have to make decisions related to crowding while uncertain about capacity or disruptions.

Thesis Research Goal III. *Develop decision strategies that can be utilized in case of incomplete information or uncertainty about travel options that can be used for decentralized decision making in agent-based simulation.*

One of the key properties of complex systems is that there is no central controller that controls all agents. As a result, simulation studies that focus on out-of-equilibrium situations occurring due to incomplete information, overcrowded vehicles and disruptions require that agents can adapt their plan individually without having complete knowledge of the state of the system or the future states of the system. The third research goal focuses on models that can be used to achieve this. We introduce a decision problem where an individual traveler needs to decide whether to wait for the next opportunity to travel with their preferred but currently unavailable travel option, or switch to an alternative travel option using only information from the published timetable. The difficult part is that the traveler does not know whether the preferred travel option becomes available again in the next minute, the next hour, or the next day.

We adopt ideas from *online optimization* (Borodin and El-Yaniv, 1998) and *robust optimization* (Ben-Tal and Nemirovski, 1998) where it is assumed that a passengers expects the worst to happen after a decision is made as well as ideas from *stochastic optimization* (Birge and Louveaux, 1997) where it is assumed a passenger has some

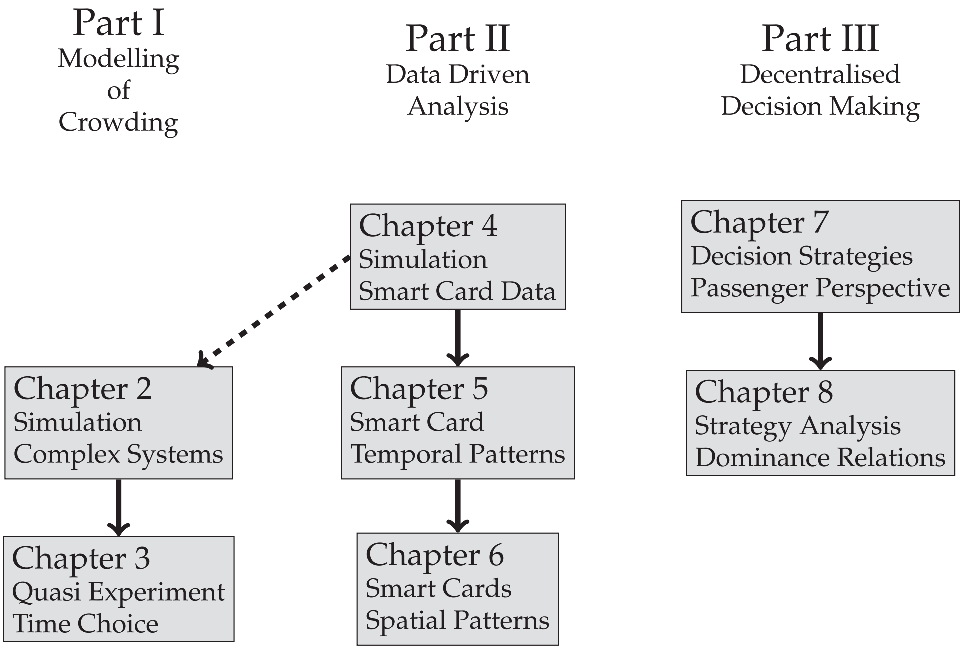


Figure 1.3: Overview of the structure of this thesis. A dashed arrow indicates that one chapter influenced another chapter, but that they can be read separately without any problems. The other arrows indicate which chapters follow each other. The parts correspond to the research goals of the thesis and the chapters are presented along with two keywords describing the topics and techniques central to the chapter.

information related to the probabilities of the different scenario's. Our analysis covers a large number of possible objectives for the passenger, which makes it possible to construct various decision rules that can be implemented by agents in an agent-based simulation.

The structure of this thesis is built around these research goals. The three parts of the thesis correspond to the three goals. In Figure 1.3 that shows the organization of the thesis and the underlying relations of the chapters, along with two important keywords for each chapter. In order to allow readers to read each chapter separately, each chapter introduces important notation and concepts separately.

1.4 Chapter Summaries

In Chapter 2 we introduce a framework inspired by the El-Farol Bar Game where we have heterogeneous agents that make repeated travel decisions while an operator makes operational decisions that influence the agents. We conduct two studies: One where we evaluate the effect of information on the agents, where we find that informing agents about the crowding of certain travel options results in higher utilization of public transport but lower satisfaction. This relationship is influenced by the ratio of potential public transport users and the capacity. In the second study we have a population of agents that have a random origin-destination along a line-network. Here we measure the impact of the way operators assess demand for the allocation of capacities in the network: Different approaches lead to different cost/benefit ratio's for the operator and different levels of public transport use.

In Chapter 3 we investigate the choices of a large group of students in a quasi-experiment about time-choice when commuting by public transport. All participants make twenty decisions without information about crowding, and another twenty decisions after a crowding indicator is introduced. Based on behavioral measures that indicate whether individuals are frequent switchers and how many different travel options they switch between, we can distinguish four types of travelers. We observe that information influences the sizes of the populations associated with these behavioral types. The behavioral types and their reaction to information provides a base for the construction of more realistic agent types in future research.

In Chapter 4 we use smart-card data to generate an agent-based simulation scenario for the software package MATSim (Horni et al., 2016). It provides a first step necessary to use smart-card data in agent-based simulations where individual travelers are modeled. To achieve this, an efficient pipeline for processing smart card data was implemented that is able to analyze for many smart cards at which station most time is spent, resulting in an educated guess for a home and a work location. Unfortunately, the default economic decision model of the agents was not suitable to perform simulation studies for revenue management. However, the processing of smart-card data and running a large-scale simulation based on it turned out to be computationally feasible.

The smart-card data processing pipeline engineered in Chapter 4 paved the way for data mining approaches that go beyond mere home and work patterns. In Chapter 5 we focus on temporal-patterns observed at different stations. We propose a methodology that processes time-intervals using multiple runs of k-means clustering per station. With this method we find that shorter time intervals, typically associated with lunch or shopping activities can be detected besides the common home-work travel patterns considered in Chapter 4.

In Chapter 6 we propose a method to generate synthetic smart card data from predefined activity patterns and propose a way to validate methodologies that detect activity patterns by comparing the outcome to the predefined patterns. Furthermore, we use the kind of temporal patterns developed in Chapter 5 to estimate the frequency of these patterns at stations to obtain a discrete distribution of temporal patterns per station. Based on these distributions it is possible to cluster the stations themselves, providing insight in how similar stations are regarding the activity of passengers around that station. Using the validation approach, we find that the proposed methodology works well for stations with a clear activity purpose, such as residential areas for home activities or office areas for work activities. Stations where multiple activity types take place are still a challenge that needs to be addressed in future research.

In the final part of the thesis, we make a first step toward modeling individual decision strategies. While these decision strategies for individual passengers to produce useful advice within a smart-phone app, they are also important if we want to model decentralized decision strategies for individual passengers in agent-based simulations. In Chapter 7 we consider how passengers can decide whether to take a detour or not in case of uncertainty. In the chapter we analyze how a passenger can decide whether to take a detour or wait at the current location in case of a disruption. Modeling the reaction of passengers in case of incomplete information is a major challenge and one of the problems considered in Van der Hurk (2015); Van der Hurk et al. (2017), which considers what advice passengers need from the operator, while taking limited capacity into account. In crowded situations, it is often uncertain at which point in time the passenger will fit in a vehicle. These proposed decision models can help to build agent-based models where the agents are able to make decision if they are unaware of all operational details, in the same way passengers are unaware of these details in real life.

In Chapter 7 we introduce a model where a passenger has arrived at the station to discover that the preferred travel option will be unavailable for an unknown amount of time. The option can either become available within minutes, or it may take very long. The passenger can decide to use an alternative, less preferred choice, or wait for the intended option to become available. Different decision strategies to deal with this situation can be formulated, even if there is no information about the length of disruption. In Chapter 7 this problem is solved computationally for a specific objective of the passenger. In Chapter 8 we consider when it makes sense to consider all strategies, or whether some of the strategies can be discarded beforehand. This is done in relation to a very general class of objectives of the passenger.

1.5 Authorship

With the exception of Chapters 3 and 8, the author of this thesis was the principal researcher of the presented work. This means that the author did all implementation, analysis and writing for these chapters. Discussions with Leo Kroon resulted in significant improvements of the research and writing in each chapter. In Chapters 2, 4, 5 and 6, discussions with Peter Vervest were invaluable in shaping the direction of the research and formulating research questions. The discussions and feedback provided by Evelien van der Hurk were vital in the development of Chapters 4, 5, 6. Gábor Maróti provided a number of good suggestions in the modeling of Chapter 2 and had essential comments for the clarity of its introduction. Ting Li kindly provided the smart card data that was used in Chapters 4 and 5. For Chapter 4 Milan Lovric helped with the implementation needed for the use of smart card data in MATSim.

Chapter 3 is the result of a collaboration with Jan van Dalen and Clint Pennings who are involved in the Statistical Methods course of the Rotterdam School of Management. They proposed to do a quasi-experiment within this course to obtain data inspired by the model introduced in Chapter 2. They both designed the quasi-experiment, while Clint Pennings implemented the software used to collect the data. The analysis and writing of the chapter were performed by the author of this thesis, with the help of valuable feedback of Jan van Dalen, Clint Pennings and Leo Kroon.

Chapters 7 and 8 are joint work with Leo Kroon, Marie Schmidt and Anita Schöbel. The problem definition and model for possible strategies was the result of joint discussions with all four authors. Chapter 7 was written by the author of this thesis with valuable comments from all co-authors. The theorems and results of Chapter 8 were developed by Marie Schmidt and Leo Kroon. Marie Schmidt performed the majority of writing for this chapter, while Anita Schöbel and the author of this thesis checked the theorems and proofs and made numerous suggestions for improving the chapter.

1.6 Conclusions and Future Research

Using an extension of the El-Farol Bar Game, we were able to model and simulate a complex system that captures crowding interactions of public transport passengers, system capacities and the public transport operator. With this model we can observe varying degrees of crowding when passenger behavior, information and/or capacity allocation policies are changed.

We have shown how smart card data can be transformed into an agent population and use that population to simulate the journeys and travel of decisions of the agents. We have proposed a number of methodologies to extract temporal and spatial pattern from the smart card data, that can be helpful in activity based modeling of transport demand. By providing a method to validate our and future methodologies, we can observe that some types of activities can be detected rather well, while activity patterns at locations with multiple activity types are a lot harder to detect and distinguish.

Finally, we developed and analyzed decision strategies that can be utilized to develop agent implementation that can make decisions with incomplete information or uncertainty about disruptions. These models are a first step toward agent implementations that are confronted with full vehicles or unexpected disruptions, without being controlled centrally by the public transport operator or a global optimization model.

With these results, we can conclude that complexity science can be used to model and analyze crowding within public transport. For future research it is interesting to integrate the separate parts in this thesis and build a realistic model that can help public transport operators deal with crowding in a way that is beneficial for passengers and reduces operational costs for the operator. By combining the crowding interactions of Part I, the temporal and spatial activity patterns of Part II and existing models developed by researchers in the field of transport economics, NS and other operators could develop a tool that can be helpful the next time they face severe capacity restrictions. In transport systems where passengers do not have information on crowding or public transport can be unreliable due to filled vehicles, the work in Part III provides extensions that can be used to include decentralized decisions of individual passengers in this integrated approach.

Part I

Crowding in Public Transport as a Complex System

2

Capacity, Information and Minority Games in Public Transport

Co-authors : Leo Kroon, Gábor Maróti and Peter Vervest
This paper has been published in Transportation Research: Part C.
The full reference is:
P.C. Bouman, Leo Kroon, Peter Vervest, Gábor Maróti, Capacity, Information and minority games in public transport, Transportation Research Part C: Emerging Technologies, Volume 70, September 2016, Pages 157-170.
Under the title *Rolling Stock Allocation and Crowd-Sensitive Passengers* it won the second place for the *best paper award* of the Rail Tokyo conference held in 2015.

2.1 Introduction

Operators in public transport are often faced with peak demands, typically during the morning and afternoon rush hours. As a result, vehicles can become very crowded, greatly reducing the comfort experienced by the passengers. The increasing emergence of information technologies creates many new opportunities for public transport operators. They can now use smart phone apps to communicate with passengers, providing advice and information on their journey. Decision support systems are influencing the way operational decisions are made. The increasing adoption of smart card technologies provides operators with a lot of new data that can be used by these decision support systems. Detailed schedules that used to require a lot of manual labor to create can be adapted to new circumstances with the push of a button in the near future, due to improved algorithms and faster computers. Passengers get more and more opportunities to work remotely via the Internet, creating greater flexibility in when they have to be at the office. They can also use apps to make travel choices in a more strategic manner, for example if they want to avoid vehicles which are too crowded.

However, the impact of crowding on passenger behavior and the interaction between railway operations and passengers is not well understood. In this chapter we develop a model, based on the concept of minority games, that allows us to study the dynamics of crowding in public transport through computational experiments and evaluate the impact of operational and behavioral models on a number of performance measures, most importantly the utilization of available capacities. We show that we can experiment and study interactions between passengers as well as the interactions between the public transport operator and the passengers. In particular, we investigate how information impacts passengers under different assumptions about the way they process information. Operators can adopt different strategies for the optimization of capacities based on data collected about passengers travel choices. The strategies can impact these travel choices, causing a feedback loop.

Since the "El-Farol Bar Game" (Arthur, 1994) was first introduced in 1994, the concept of the *minority game* has received a lot of attention from researchers. One of the great strengths of this model lies in the simplicity of its description: There is a population of decision makers who have to decide every Thursday night whether to go to the bar. Once they go the bar, they have a positive payoff if less than 60% of the population goes to the bar, while they have a negative payoff if it is too crowded. As everyone makes this choice every Thursday, the El-Farol Bar Game has an iterative nature. While historic information is provided, the interesting aspect comes from the fact that there is no direct coordination between the decision makers.

Issues related to limited availability of resources and a lack of explicit coordination occur in many real world systems. The applications of these models include car traffic (Bazzan et al., 2000), congestion in computer networks (Huberman and Lukose, 1997) and financial markets (Challet et al., 2004). While these types of applications were considered earlier from a game theory perspective, most notably under the name of *congestion games* (Rosenthal, 1973), the novelty from the "El-Farol Bar" study was the application of a complex systems approach enabled by simulation of a repeated game, while game theory is mostly concerned with the properties of equilibria.

In this chapter, we focus on minority games where the operator cannot control agent behavior, but has control over the disclosure of information and the system capacities. The main application domain is public transport systems, where passengers share vehicles depending on their chosen route and time of travel. If a connection is operated frequently, passengers with some flexibility in their schedule can try to avoid crowded situations by shifting their time of travel. Since it is reasonable that a passenger does not want to travel at any time, we introduce the concept of individual choice sets representing the acceptable choices. To our best knowledge, this type of heterogeneity of the choice sets has not been studied in the context of minority games before.

Within public transport systems new technologies have introduced many opportunities to provide passengers with additional information: Many stations and vehicles have screens with travel information, and many passengers use smart phones to receive information during their journeys. The increasing adoption of smart card ticketing systems allows operators to have accurate data on the utilization of each vehicle. As operators in railway and metro systems can extend or shorten the trains (Fioole et al., 2006) and bus operators can employ different vehicle sizes, new operational models that include adaptive capacity allocation based on (almost) real-time information are becoming a possibility.

The main observation in the original "El-Farol Bar Game" simulations (Arthur, 1994) is that even though individual decision makers keep switching their preferred predictive model from an individual fixed set of random models, the average utilization of the bar converges to the efficient level. In order to explain this phenomenon the minority game was introduced, where the utilization history was replaced with a history of binary values indicating whether the bar was overcrowded or not. The main idea of this approach is that the set of all possible deterministic strategies can be finitely characterized so that methods from statistical mechanics can be applied (Challet et al., 2000).

In the scientific literature of different different fields different names are used for entities which make decisions. In the game theory literature, these entities are often called *players*. In the Operations Research literature these entities are often referred to by the name of their role in the operational process, for example 'dispatcher', 'manager' or 'operator'. In settings where simulation is used and many autonomous decision makers interact, the term *agent* is popular for these decision making entities. In this chapter we use the term *agent* to refer to modeling entities that represent passengers, and use *operator* to refer to the public transport operator (which can be regarded as a very special type of agent).

The remainder of this chapter is organized as follows: In Section 2.2 we introduce our class of minority games. In Section 2.3 we discuss the architecture of the components of our simulation. This simulation framework is then applied in order to investigate the effect of different information policies in Section 2.4. In a second simulation study we evaluated the effect of rolling stock optimization in the context of public transport (Section 2.5). In Section 2.6 we show that the inclusion of individual choice sets and scoring functions leads to NP-hardness of maximizing the efficiency of a given system. We discuss our findings and plans for future research in Section 2.7.

Related Work

A variation of the minority games are the *resource allocation games*, introduced by (Galstyan et al., 2003). This extension of minority games introduces multiple resources and capacities that vary over time. Conditions are given under which the agents can use a social network structure in order to adapt efficiently to variations of the capacities. The fluctuations of the capacities considered in the studies associated with the *resource allocation games* only depend on time and do not depend on the distribution of agents over the resources during the game.

While the body of knowledge on learning techniques for agents in minority games (Kets, 2012) is very useful for the engineering and design of artificial agents, it is a question whether it is applicable within systems where real humans are involved. Selten et al. (2007) conducted a laboratory experiment involving route-choice. The participants could be divided into three groups: Participants who had the tendency to switch away from a road if it was congested during the previous round, participants who had the tendency to stay on their current road regardless of it being congested during the previous round, and participants who were harder to classify. Although the participants showed different types of behavior, the distribution of the participants over the roads approached the equilibrium very closely.

The problem of congestion in transportation networks has been addressed by researchers using *agent-based models*. One example is the research by Kaddoura et al. (2015). It studies a setting where a bus operator has to choose fares and headways and has to compete with possibly congested road traffic. An external search process is used to test different combinations of fares and headways and for every combination the MATSim simulator (Horni et al., 2016) is used to evaluate the travel equilibrium. This way the impact of departure time choice on the operator, the passengers and the social welfare can be analyzed. In a second study due to Zou et al. (2015), an agent-based choice model is developed, where decision rules based on data collected using a survey. This survey asks for the last trip and performs a stated-choice experiment. these decision rules are combined with a knowledge learning process in order to perform a case study which involves the subway and car traffic in an area of Beijing. They find different spreading behavior of peak traffic when the number of travelers is increased and for different congestion charges which road traffic has to pay.

2.2 A Model For Crowding Dynamics

The general scheme of our model is that in each round every agent decides whether it will use one or more resources or refrains from doing so. Using a resource gives the possibility to gain a positive payoff or a negative payoff depending on the utilizations encountered. If it does not use any resource, the payoff will be neutral, i.e. zero.

We define symbols for the resources, the agents and payoffs. The resources are defined in the following way:

- A set $\mathcal{M} := \{1, 2, \ldots, m\}$ of m resources.
- A *soft* capacity function $\text{cap} : \mathcal{M} \to \mathbb{Z}^+$.

Thus there are m resources, each of which having an associated capacity. Note that we define *soft* capacities that can be violated, but everyone in such a situation should have a negative payoff. Based on the capacity we define the utilization of a resource as the fraction of its capacity that is occupied. The typical example in public transport is the number of passengers divided by the number of seats. As the game is played iteratively, the transport operator can adapt the capacities based on observations recorded during earlier rounds of the game. We also define the preferences and payoffs of the agents that play the game:

- A set $\mathcal{N} = \{1, 2, \ldots, n\}$ of n agents.
- A non-empty collection C_i of subsets of $\mathcal{M}$.
- A scoring function $s_i : \mathbb{Q}^m \times C_i \to \mathbb{R}$ for each agent $i \in \mathcal{N}$

During each round, every agent should choose one of the options in its choice set. We assume that every choice set contains the empty set as a neutral option, but this is not strictly necessary. We can describe the outcome of a round based on the choices made by all agents. If an agent i chooses to use a set of resources $c \in C_i$, we set the indicator variable x_{ic} to 1. The set of all vectors of x_{ic}'s describing a valid outcome is thus defined by

$$\mathcal{O} = \{x \,|\, \forall i \in \mathcal{N} : \sum_{c \in C_i} x_{ic} = 1, x_{ic} \in \{0, 1\}\}. \tag{2.1}$$

Given the outcome vector x for a round, we can calculate the utilization of the resources. We define a vector $u(x) \in \mathbb{Q}^m$ that contains an entry for each resource. The entry $u_r(x)$ for resource $r \in \mathcal{M}$ is calculated as follows:

$$u_r(x) = \frac{\sum_{i \in \mathcal{N}} \sum_{c \in C_i : r \in c} x_{ic}}{\mathrm{cap}(r)}. \tag{2.2}$$

While in principle s_i can be a general scoring function, for ease of analysis we use the restricted class of *threshold based scoring functions*. These scoring functions have a payoff of -1, 1 or 0 depending on an individual threshold θ_i and the maximum encountered utilization. The scoring function itself is then defined as follows:

$$s_i(u, c) = \begin{cases} 0 & \text{if } c = \emptyset, \\ 1 & \text{if } \max_{r \in c} u_r \leqslant \theta_i, \\ -1 & \text{otherwise.} \end{cases} \tag{2.3}$$

Performance Measures

Since we want to analyze the behavior of an agent population, we introduce some measures that are of analytic interest and can be recorded during a simulation. We define the # symbol to denote the cardinality of a set (e.g. $\#\{6, 9\} = 2$). Given an outcome $x \in \mathcal{O}$ during any of the rounds of the game, we can calculate the following observations:

- $\text{utl}(x) = \frac{1}{n}\#\{i \in \mathcal{N} : x_{ic} = 1, c \neq \emptyset\}$, i.e. the fraction of agents utilizing any resource (i.e. the fraction of those not selecting the empty set).

- $\text{pos}(x) = \frac{1}{n}\#\{i \in \mathcal{N} : x_{ic} = 1, s_i(u(x), c) > 0\}$, i.e. the fraction of agents with a positive payoff.

- $\text{posc}(x) = \frac{\text{pos}(x)}{\text{utl}(x)}$ is the fraction of agents with a positive payoff among the agents who utilize a resource.

- $\text{avg}(x) = \frac{1}{n} \sum_{i \in \mathcal{N}} \sum_{c \in C_i} s_i(u(x), c) x_{ic}$ is the average payoff of the agents.

In real world public transport systems, operators can gain insight into the number of travelers, which is similar to the utl measure, using data collected by their smart-card ticketing systems. The payoff of travelers is harder to measure for real travelers, but operators can include features in their smart phone apps that ask for and record the satisfaction of passengers.

2.3 Architecture of the Agents and Simulation

Given an instance of the game, a simulation still depends on two more aspects. (1) The way the agents make their decisions and (2) to which extent the agents can observe the outcome of the previous rounds. As we want to be able to evaluate the effect of different types of agent behavior, we allow different types of agents in the population. We introduce these types in Section 2.3.1. We first define the main steps that are executed in each round of the simulation:

1. Let every agent $i \in \mathcal{N}$ choose one option $c \in C_i$ from its choice set according to its agent type.

2. Calculate the outcome vector x and corresponding utilization vector $u(x)$ accordingly.

3. Let every agent $i \in \mathcal{N}$ observe, learn and process its score $s_i(u(x), c)$ based on its agent type.

4. The operator lets every agent $i \in \mathcal{N}$ observe, learn and process information based on the active information policy and the utilization vector u.

From these steps we can see the necessary ingredients for an agent implementation within this simulation scheme. An agent needs a choice function and can optionally implement a method to process incoming scores and information.

2.3.1 Agent Types

The outcome of the simulation depends to a large extend on the strategies utilized by the agents. If there would be a single deterministic strategy used by the agents, the distribution of the agents over different travel options would solely depend on their choice sets and likely lead to empty or very crowded travel options. In order to let agents distribute themselves evenly over the different travel options without explicit coordination it is useful to let them include some random experimentation in their decision process. There are many ways to include randomness in the decision behavior. We have chosen to use the following *agent types* in our experiments, as we believe they are simple to understand and/or close to the decision rules used in many agent-based models.

The most straightforward agent type is the *random agent*, who selects a choice from its choice set uniformly at random in each round. This agent type is useful for both benchmarking purposes, validating the simulation architecture analytically, and to model noisy behavior within the population.

The more complicated agent types make decisions based on observations during earlier rounds of the game. For these agent types, step 4 of the simulation process in a round can have an effect on step 1 in the next round. The number of rounds the agents look back is referred to as the *memory length*. An important finding in the minority game model is that the most efficient utilization is reached when the memory length of the agents is proportional to the logarithm of the total number of agents (Savit et al., 1999).

The second type of agent, the *average payoff agent*, applies a simple reinforcement learning heuristic. Reinforcement learning strategies have received notable attention in the literature, and we take one of the most simple ones as an example. As such, our *average payoff* agents perform exploration during 10% of the rounds by making a random choice, while they exploit the best option based on observed average payoffs during 90% of the rounds. In case multiple choices have the best average payoff, the tie is broken by picking one option uniformly at random.

A variation of the average payoff agent is the *average utilization agent*, who uses the same reinforcement learning heuristic to learn the average utilization of the resources. The main difference is that this agent uses the information received to learn the average utilization and picks the choice with the lowest average utilization, or the neutral option if this choice has still higher average utilization than its threshold.

The last type of agent, the *predictive agent*, aims to predict future utilizations in order to find the best choice. If the agent can predict future utilizations, the agent can generate a fictitious utilization vector and evaluate the expected score of each choice. This agent type is similar to the one studied in the original El Farol Bar paper (Arthur, 1994). In a round with index t, the agent checks which of its personal heuristics was most accurate in round $t-1$ and uses this one to predict utilizations in round t. As our model introduces the concept of multiple resources, there can be situations where an agent does not know all historic utilizations of each resource. We calculate the accuracy of each heuristic based only on the information that is available. If there is no information about round $t-1$ for a particular resource, the agent will consider the information of the most recent time-step for which information is available. As availability of utilization information is defined on the agent level, an agent can compare the heuristics using the same data set.

We implemented the following predictive heuristics: Replicate the oldest utilization in memory, take the average of the utilizations in memory or fit a simple linear regressive model on the utilizations in memory.

2.3.2 Information Policies

At the end of each simulation round, we let each agent process information and observations on the utilization of the resources. We define a unit of information as a 3-tuple (t, r, u) consisting of the round of the game t, a resource r and a utilization vector u. As the agent can have multiple resources in its choice set, it should be able to receive and process multiple pieces of information each round. In general, an *information policy* is a set of rules that determine the information offered to each agent in each round. While there are very many information policies possible, we propose four basic ones.

In public transport, the fact that an agent is using a resource allows it to observe the utilization. Thus in our most basic information policy, *private information*, an agent receives exact information for the resources related to its most recent choice.

On top of private information, the entity or agent controlling the resources could monitor the utilizations and try to attract more agents in case a resource r has a low utilization, say less than 40% of its capacity. In such a situation the information policy can state that additional information regarding a resource r should be provided to all agents. We refer to this type of policy as *adaptive information*.

In some situations there are information systems that provide information on the crowding of a resource. A real life example one can think of is a smart phone application of a public transport operator, that shows one, two or three icons based on the forecasted crowdedness of a vehicle, reducing the utilization level provided to the agent to a few discrete values. This idea is captured by the *estimate information policy*, where the utilization of each resource is rounded up to either 0, $\frac{1}{4}$, $\frac{2}{4}$, and so on, similar to the 3 symbol crowding indicators provided by some operators. This rounded utilization is then provided to all agents. We should take care that we send the rounded utilizations in case the agent did not observe the utilization, and use exact utilization otherwise.

In the final template, we send out exact information on every resource to every agent in each round – thus in this situation the agents have *full information*.

All these information policies can be interpreted in terms of smart phone applications, except the *private* policy where no information is communicated. The *adaptive* policy can be interpreted as an app that gives a notification in case a train is not utilized a lot, the *estimate* policy can be interpreted as an app that shows up to four symbols after a possible journey to indicate the crowding, and the *full* information policy can be seen as an app that shows passenger numbers and capacities for each train.

2.4 Evaluating Information Policies

In our first experiment, we evaluate the four information policies in a population of agents that use public transport to travel from a single origin to a single destination, but can choose different times of travel. As such their choice sets contain only singleton resources, reflecting the departure times a public transport service is scheduled and the empty set as a neutral option, reflecting a journey by car or staying at home. We find that increasing the available information leads to a greater number of agents utilizing the public transport system, but at the cost of the average payoff. However, the magnitude of this effect is influenced by the ratio of population size and available capacity.

2.4.1 Experimental Setup

In our experiments, we work with $m = 10$ resources representing the departure times. Every choice set C_i contains $\emptyset$ and 3 different singleton sets picked uniformly at random from $\mathcal{M}$ without replacement. For each agent i we use a threshold based scoring function with $\theta_i \in \{\frac{5}{10}, \frac{6}{10}, \ldots, 1\}$ picked uniformly. The capacity of each resource is fixed to 10, i.e. $\mathrm{cap}(r) = 10$.

As we have 100 units of capacity available each round, we consider a high capacity scenario with $n = 50$ agents, a regular scenario with $n = 100$ agents and a low capacity scenario with $n = 200$ agents. For each of these scenarios, we vary the population by picking all pairs $p, q \in \{0, 1, \ldots, 10\}$ such that $p + q \leqslant 10$. Our population then consists of $10p$ *random* agents, $10q$ *average utilization* agents and $10(10 - p - q)$ *predictive* agents. In total 66 population mixtures are evaluated. For each mixture of agent implementations we regenerate the choice sets and thresholds 100 times. For a given instance of the choice sets we run the experiment 25 times, regenerating the predictive agents 5 times if they are part of the population. Thus, in total we run 2500 simulations per combination of population mixture and population size. As we want to ignore the warm-up period of the simulation and like to interpret the rounds as days, the measures are recorded from round 10 to 40 during each simulation run.

The predictive agents each have an individual randomly selected set of 3 random predictive heuristics from the following list: Average heuristic with memory lengths of either 4, 5, 6 or unlimited, linear regression with memory lengths of 4, 5, 6 or unlimited, replicate the oldest observation with memory length either 1, 2 or 3.

Table 2.1: Results of the simulation study where different information policies are evaluated. The minimum, average and maximum utl (fraction of agents utilizing a resource) and posc (fraction of agents who have a positive payoff among those that utilize a resource) values measured for each of the 66 population mixtures are reported.

(a) Results for utl

	Private	Adaptive	Estimate	Full
utl ($n = 50$)				
Minimum	0.63	0.75	0.75	0.75
Average	0.77	0.90	0.91	0.91
Maximum	0.91	0.97	0.99	0.99
utl ($n = 100$)				
Minimum	0.45	0.66	0.70	0.73
Average	0.60	0.70	0.73	0.77
Maximum	0.75	0.75	0.82	0.85
utl ($n = 200$)				
Minimum	0.20	0.40	0.46	0.51
Average	0.42	0.48	0.52	0.56
Maximum	0.75	0.75	0.75	0.78

(b) Results for posc

	Private	Adaptive	Estimate	Full
posc($n = 50$)				
Minimum	0.90	0.90	0.89	0.87
Average	0.95	0.93	0.91	0.90
Maximum	0.99	0.97	0.93	0.92
posc ($n = 100$)				
Minimum	0.46	0.46	0.41	0.30
Average	0.77	0.65	0.60	0.48
Maximum	0.98	0.84	0.72	0.54
posc ($n = 200$)				
Minimum	0.02	0.02	0.02	0.02
Average	0.49	0.35	0.27	0.19
Maximum	0.94	0.68	0.42	0.26

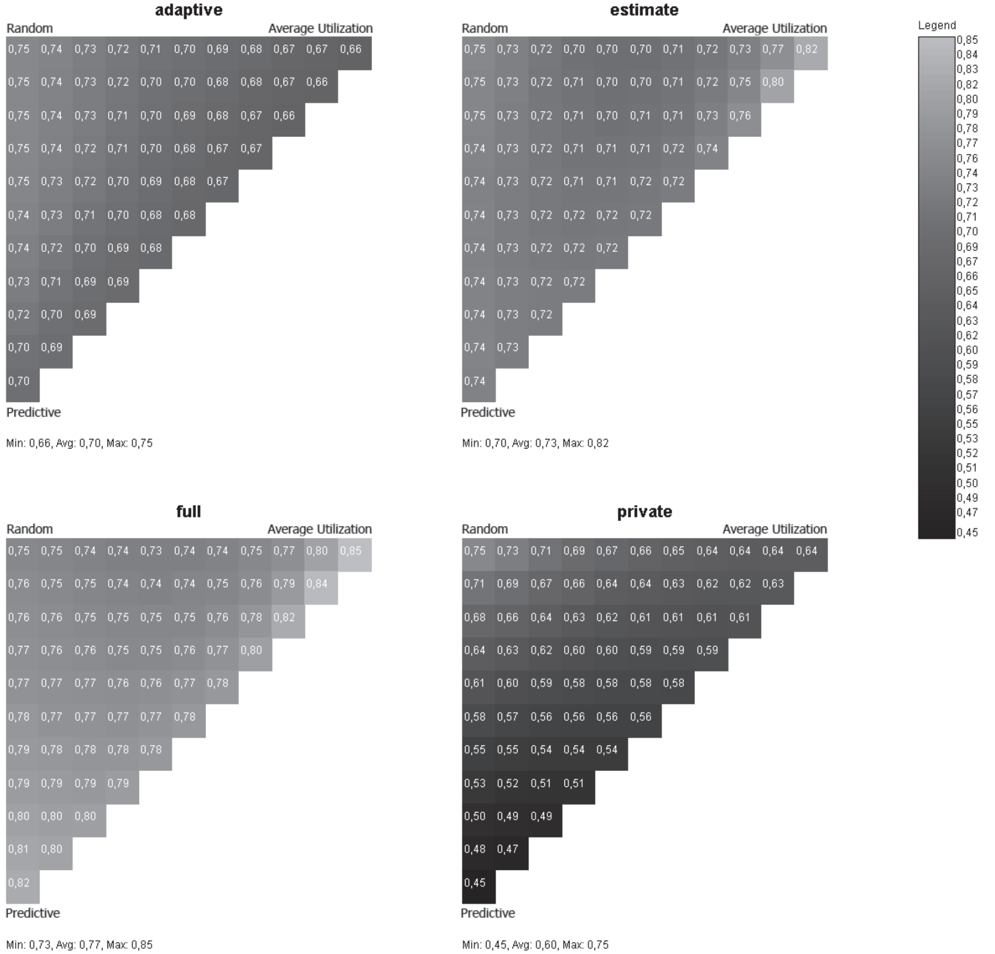


Figure 2.1: The effect of different mixtures of agent types on the utl measure for $n = 100$. The corners of the triangles correspond to a population with a single agent type, while other points in the triangle correspond with mixtures.

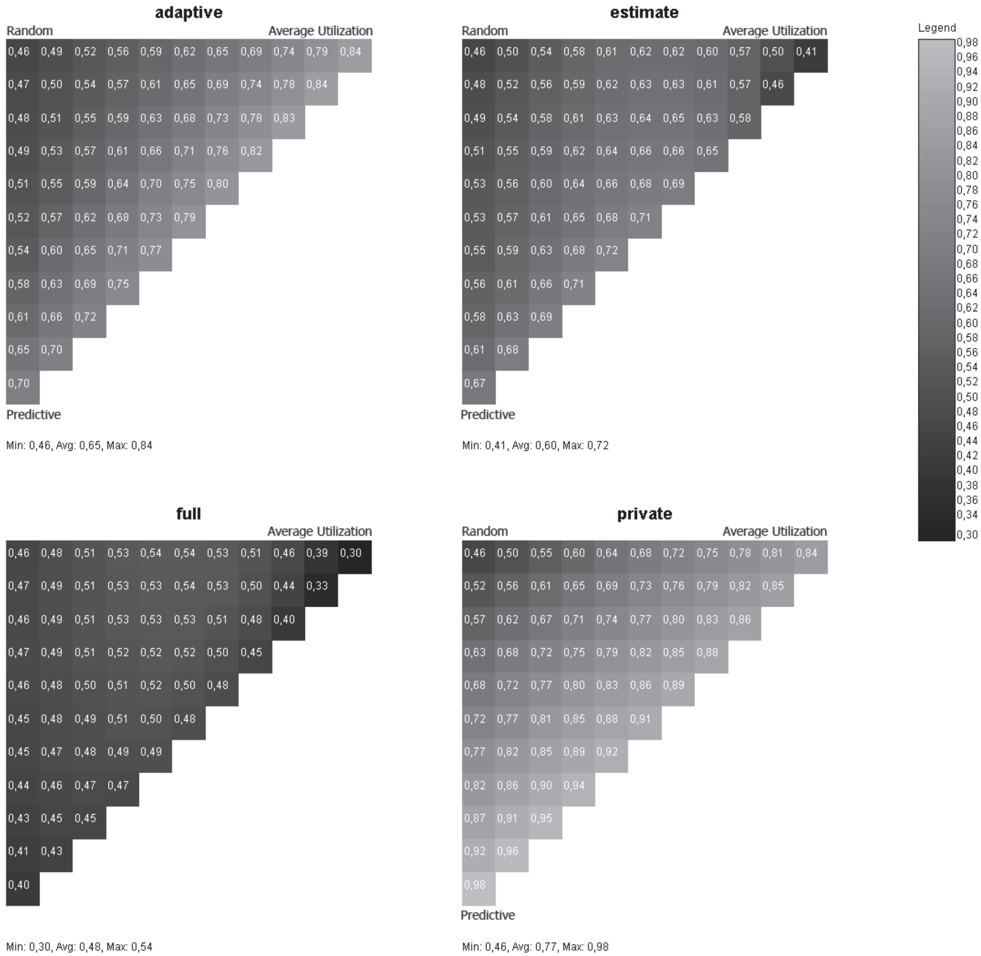


Figure 2.2: The effect of different mixtures of agent types on the posc measure for $n = 100$. The corners of the triangles correspond to a population with a single agent type, while other points in the triangle correspond with mixtures.

2.4.2 Results and Discussion

The results of our simulation experiments are presented in Table 2.1. If we look at the left column of Table 2.1, we can verify that when we increase the level of information provided to the agents, the number of agents utilizing a resource increases. If we look at the average values from private to adaptive, we can see that the 0.13 increase for the $n = 50$ scenario is greater than the 0.06 increase for $n = 200$. These numbers suggest that the effect of information depends on the units of capacity available per agent in the population.

If we look at the fraction of agents utilizing a resource with a positive payoff (this can be interpreted as customer satisfaction) in the bottom half of Table 2.1, we can see that increasing the level of information decreases the posc value. This seems intuitive, since adding information attracts more agents, and having more agents increases the likelihood of crowding. Again, the amount to which the posc value decreases when we move from the *private* to the *adaptive* case is impacted by the amount of capacity available per agent: For $n = 50$ the decrease of 0.02 is less dramatic than the 0.14 decrease in the $n = 200$ case.

We can use Figures 2.1 and 2.2 to study the effect of the different agent types. As expected, information has no impact when all agents decide *randomly*, ignoring information observed in the past. When at least some of the agents use information to make their decisions this has an influence on the observed utl and posc metrics. The way information is processed also influences the observed measures: A population consisting of only *average utilization* agents has a different reaction than a population of *predictive* agents. This suggests that the decision strategy of the agents is an important aspect to consider for applications of this model.

The impact of disclosing more information to the agents starting at *private* information and moving to *adaptive* information is highest for the *predictive* agents (utl goes from 0.45 to 0.7), while the effect for the *average utilization* agents is greatest when moving from *adaptive* information to *estimate* information (utl goes from 0.66 to 0.82). As a result a good choice for a suitable information policy depends on the way the agents process the information.

For future work we will investigate whether better information policies can be designed. There are also questions regarding the effect of noise in communications, such as technical problems at the side of the operator or agents ignoring information sometimes. We are curious to learn whether such noise could lead to less correlated agent behavior and whether this can lead to better system efficiency. Another aspect that plays an important role is the *memory* of the agents: To what extend does old information contribute to new decisions. The *average utilization* agents average over the complete history, so it seems likely that new information has little influence on their behavior after a large number of rounds, while the *predictive* agents can have shorter memories, which should have an influence on whether the system converges to a stable state or keeps fluctuating after a large number of rounds.

2.5 Capacity Optimization in Public Transport

In our second experiment, we want to evaluate the effect of rescheduling capacities on the crowding dynamics in the system based on predictions of demand the operator makes using historical utilization of vehicles. In practice these types of predictions can be based on detailed smart-card data (Pelletier et al., 2011). Consider a public transport scenario where a train moves back and forth a railway line with 5 stops. The train drives 8 full cycles per day and as moving along the railway line in one direction gives us 4 trips between the stops, the timetable consists of $4 \cdot 2 \cdot 8 = 64$ trips that are offered each day. As individual travelers want to travel between two stops that are not necessarily connected by a single trip, a journey can consist of one or multiple trips. We assume passengers will travel directly toward their destination and as a result for each origin-destination (OD) pair there are 8 different time slots at which passengers can make their journeys.

In order to facilitate the flow of passengers, the trains need to be long enough in order to allow comfortable transportation. To achieve this, the operator monitors utilization of vehicles and adapts the assigned number of rolling stock units to each train accordingly. The operator can decide how often the observed utilizations are evaluated to build a new rolling stock schedule. In this experiment, we assume that this will happen periodically. We refer to the number of rounds after which the operator produces a new rolling stock schedule as the *rescheduling period*, denoted by an integer k.

2.5.1 Capacity Allocation

As the use of rolling stock units determines a significant amount of the operational costs of a public transport operator, an operator monitors the utilization of the train vehicles and adapts the capacities if necessary. The typical model used to determine the rolling stock allocation in these situations is by constructing the network of possible train movements, specifying a minimum demand on the arcs that correspond to passenger trips and look for a circulation of minimum cost (Ford and Fulkerson, 1962) based on operational costs.

We implemented a module in the simulation that represents an operator which dynamically monitors demand. During each round of the simulation, train utilization for each trip is recorded. After k rounds, the demand of a trip is set to $\mu + 2\sigma$, where μ is the mean utilization and σ its standard deviation during those k rounds. Although public transport operators can use very advanced forecasting systems based on multiple data sources, such as the weather, event calendars and historical smart card data, we chose this rule because simple rules are easier to analyze and similar rules of thumb are employed by real operators. The capacities of each trip are then calculated according to a rolling stock circulation, where we define a cost of 1000 per unit used and a cost of 1 for moving a unit between consecutive stops on the line. These numbers represent that buying and maintaining rolling stock units is

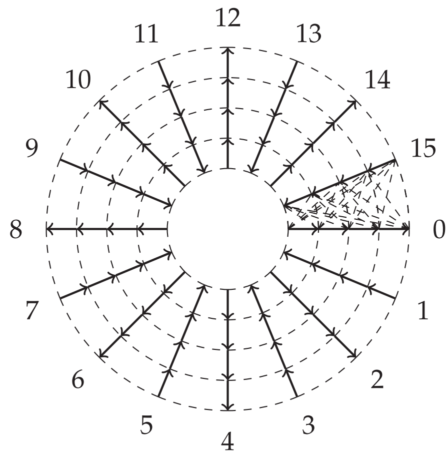


Figure 2.3: The network of trips along a line during the 16 time slots with the possible movements of rolling stock units outside the regular trips represented by dashed lines.

a lot more costly than moving them around. We also assume that storing a unit at a station does not impose any costs. As a result, the minimum cost rolling stock circulation minimizes the number of units required before minimizing the movement costs, given that the defined demand must be met.

We use a minimum cost circulation algorithm (Ford and Fulkerson, 1962) (which shares quite a lot of similarity with the well known augmenting path methods for max flow) to obtain the capacities. Although more efficient algorithms exist for this problem, the augmenting path method is straightforward to implement and fast enough for our simulations. The input network is visualized in Figure 2.3. In this representation, straight arcs represent movements between the stops and must carry the determined demands. The circular arcs represent storing a vehicle at a stop. The overnight arcs represent the purchase costs of the vehicles and the overnight balancing movements.

While the algorithms employed by operators need to take many different types of rolling stock and regulations into account (Fioole et al., 2006), for reasons of simplicity and interpretability we assume that we have only one type of rolling stock with a nominal capacity of 10 seats.

2.5.2 Experimental Setup

In order to set up the simulation, we define a resource set that consists of the trips, so based on the 5 stops and 16 time-slots, we get $m = 64$ resources. The choice set of an individual agent is generated as follows: We pick two stops $o \neq d$ from among the five stops. By choosing o as the origin and d as the destination, the direction along the railway line is defined. We then pick 3 from the 8 available time slots corresponding to this direction in order to define the acceptable journeys. The choice set then consists of the empty set and the sets of trips corresponding to the journeys drawn randomly. Again we work with threshold based scoring functions where the threshold is picked uniformly from $\{\frac{5}{10}, \frac{6}{10}, \ldots, 1\}$.

For the purpose of simplicity, we use only one type of agent during this experiment: The *average payoff agent*. One of the reasons to choose this agent implementation is that MATSim (Horni et al., 2016) takes a related approach when computing dynamic traffic equilibrium. We pick the number of agents as $n = 1000$. The reason to take a large agent population is because the available capacity is at least 640 due to the fact that every trip has at least 10 seats and we have 64 trips in total. In order to have a high probability to facilitate all the demand during the first rounds of the simulation, we set the initial demand of rolling stock units for each trip to 5. We also experimented with initial rolling stock counts of 1 and 10 units. We found that, keeping all other parameters equal, this did only affect the observed values during the first few rounds of the simulation.

Our goal is to evaluate the effect of different rescheduling periods. As the observed demand depends on the length of the rescheduling period, we evaluated rescheduling periods of 1 round, 5 rounds and 10 rounds. For each of these rescheduling periods, we generate 50 agent populations of choice sets with random thresholds. For each population we then run 2 simulations of 50 rounds.

2.5.3 Results and Discussion

The time series distributions of the observed values of utl and posc during the 100 simulation runs for each of the three policies are presented in Figure 2.4. The distributions of average payoffs, the operational costs (according to the minimum cost rolling stock circulation) and the number of rolling stock units utilized during the last round of the simulation are presented in Table 2.3.

As we increase the length of the period, we can observe that utl increases. If the rescheduling period is 1 round, we can observe from Figure 2.4a that it converges to a mean of 0.42. For a rescheduling period of 5 rounds it converges to a mean of 0.6 (Figure 2.4c) and for a rescheduling period of 10 rounds it converges to a mean of 0.66 (Figure 2.4e). One possible explanation for the fact that longer periods give a higher value for utl is that a longer rescheduling period has the potential to yield a more stable mean and possibly more accurate standard deviation (except for the case where the period is 1; then the standard deviation is always 0). As the mean and standard deviation have direct effects on the demand and thus on the capacities that are calculated, they seem likely causes for the observed behavior.

For the fraction of agents that utilize a resource and have a positive payoff, i.e. the measure posc, we can observe that it converges to a value of 0.73 for a rescheduling

Table 2.3: Results of the simulation study where the effect of rolling stock optimization on average payoff, operator costs and rolling stock units required is evaluated. The measures at round 100 of each simulation are reported, for different rescheduling periods (k) of 1, 5 and 10.

	Min.	Mean	Std.	Max.
k = 1				
utl	0.38	0.42	0.02	0.46
posc	0.60	0.73	0.05	0.84
avg	0.08	0.19	0.04	0.28
Cost	2092	2532	498.5	3112
Units	2	2.43	0.497	3
k = 5				
utl	0.58	0.60	0.015	0.66
posc	0.74	0.85	0.03	0.92
avg	0.28	0.43	0.04	0.50
Cost	3166	4120	314	4120
Units	3	3.9	0.31	5
k = 10				
utl	0.62	0.66	0.01	0.70
posc	0.76	0.86	0.03	0.92
avg	0.36	0.48	0.04	0.56
Cost	4188	4219	142	5208
Units	4	4.02	0.14	5

period of 1 round (Figure 2.4b), to a value of 0.85 for a rescheduling period of 5 rounds (Figure 2.4d) and to a value of 0.86 for a rescheduling period of 10 rounds (Figure 2.4f). The 1 round scenario has a higher standard deviation of outcomes than the other scenarios.

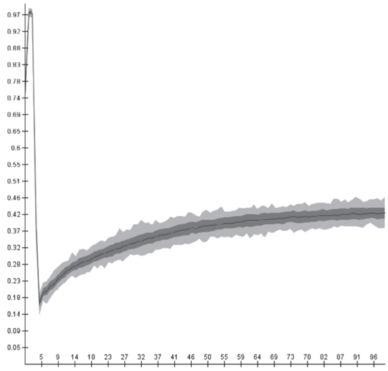
While both the utl and posc measures have higher averages for longer rescheduling periods, Figure 2.4 also shows slower convergence for longer rescheduling periods. Table 2.3 also suggests that longer rescheduling periods lead to higher costs and a higher number of rolling stock units required. However, this can be explained by the increase of the utl value. A final interesting observation in Table 2.3 is that the average payoff for the agents also increases if we use longer periods.

We can also consider the average cost per utilizing agent. For the $k = 1$ scenario this is $2532/0.42 \approx 6029$, for the $k = 5$ scenario $4120/0.6 \approx 6867$ and for the $k = 10$ scenario $4219/0.66 \approx 6256$. This suggests that the maximum profit for the operator can be obtained in the scenario where fewer agents travel by train. A similar pattern emerges when we consider the customer satisfaction measure posc, but a different pattern emerges when we consider avg. For the $k = 1$ scenario we find the average cost per unit of payoff to be $2532/0.19 \approx 13326$, in the $k = 5$ scenario to be $4120/0.43 \approx 9581$ and in the $k = 10$ scenario to be $4219/0.48 \approx 8789$. This suggests that the $k = 10$ scenario is most desirable from a societal perspective, as it minimizes the cost per unit of payoff, while the $k = 1$ scenario is most desirable for an operator purely focusing on profit. Although these observation depend heavily on the chosen cost structure in the model, this kind of cost-benefit analysis is an interesting application of the model.

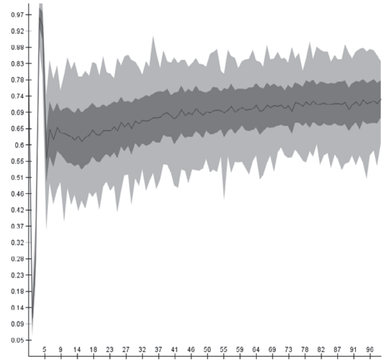
Our results suggest that there are many disadvantages for the single round rescheduling period. Increasing the period may lead to higher costs, but the number of passengers using one of the trains increases as well, which can lead to extra revenue. The only advantage of a short rescheduling period is that the costs per passenger are minimized, but from a societal perspective the cost per unit of utility is the highest. For future research we aim to search for different approaches to determine the demand for the rolling stock circulation based on the utilizations observed in the simulation. A different approach to the $\mu + 2\sigma$ rule would be to adapt demand based on observed scores. Sensitivity of the cost-benefit ratios for the relative cost between movement of rolling stock units and utilizing rolling stock units is another interesting area for further research.

2.6 Combinatorial Aspects

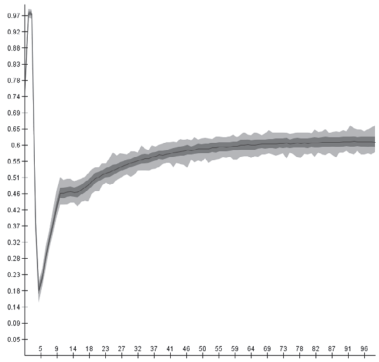
In the original "El-Farol Bar" model, it is not difficult to see that the ideal utilization of the bar lies at 60%, because all agents have the same payoff. In our extension it is not easy to determine the ideal utilization, as we are allowed to have agents with different scoring functions assigned to the same resource. As a result, it can be the case that for a single resource, some agents have a positive payoff and others have a negative one. The individual choice sets complicate matters even further. As a



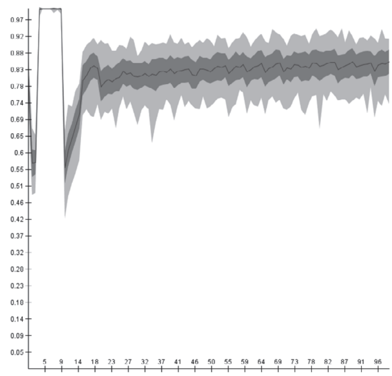
(a) utl when rescheduling period $k = 1$



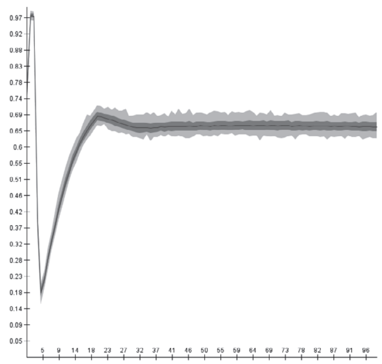
(b) posc when rescheduling period $k = 1$



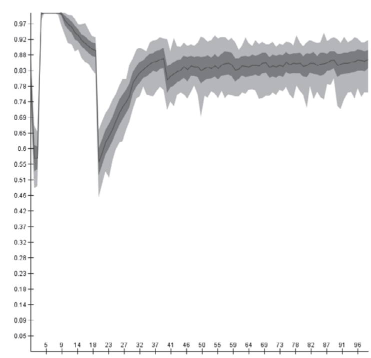
(c) utl when rescheduling period $k = 5$



(d) posc when rescheduling period $k = 5$



(e) utl when rescheduling period $k = 10$



(f) posc when rescheduling period $k = 10$

Figure 2.4: Results of the capacity rescheduling experiments. The dark line shows the mean over all 100 experiments, the dark gray area is one standard deviation away from the mean and the light gray area shows the minimum and maximum values observed.

result, it is a combinatorial problem to maximize $\text{pos}(x)$. We show this by proving the NP-completeness of the related decision problem.

Theorem 2.1. *For a given instance of the game, deciding whether there exists a valid outcome $x \in \mathcal{O}$ such that all agents have a positive payoff (i.e. whether $\text{pos}(x)$ is equal to 1) is NP-complete, even if we have threshold scoring functions with 2 different thresholds and we allow only singleton resources in the choice sets.*

Proof. We show NP-hardness by reduction from the k–SET COVER problem (Karp, 1972). In the k–SET COVER problem we are given a collection $\mathcal{A} = \{A_1, \ldots, A_n\}$ of n sets, a set of all elements $U = \bigcup_{i \in N} A_i$ and a positive integer k. We have to decide whether there exists a subset $\mathcal{A}' \subseteq \mathcal{A}$ such that $|\mathcal{A}'| \leqslant k$ and $\bigcup_{A \in \mathcal{A}'} A = U$.

We now introduce $|U|$ regular agents and $|\mathcal{A}| - k$ grumpy agents. We introduce a mapping between the sets in $\mathcal{A}$ and the resources. Each element $e \in U$ is represented by a regular agent which has a choice set that consists of singleton resources corresponding to the sets in $\mathcal{A}$ containing e. The grumpy agents have a choice set with a singleton for every resource. We define the payoff functions such that the regular agents have a positive payoff as long as they have chosen a resource, and the grumpy agents have a positive payoff if they are exclusively assigned to a resource (if we fix all $\text{cap}(r) = 1$, then $\theta_i = n$ if i is a regular agent and $\theta_i = 1$ if i is a grumpy agent). As a result the grumpy agents can only have a positive payoff if they are assigned to resources in such a way that all the other agents can be assigned to the remaining resources. By construction of the choice sets, this is only possible if the remaining k resources that are not utilized by the grumpy agents correspond to sets that are able to cover all elements. Thus, we have reduced the k-SET COVER problem into our decision problem with 2 threshold scoring functions and singleton choice sets.

NP-completeness then follows from the fact that given a vector x, we can easily check whether it is feasible and whether indeed $\text{pos}(x) = 1$. $\square$

In order to understand how the reduction works, we provide an example in Table 2.4. Here the A's and e's represent the sets and elements of the k-SET COVER instance. The corresponding model instance contains the a agents for the elements and $5 - k = 2$ grumpy agents denoted by the g agents. For $k = 3$, we can assign the two grumpy agents to resource 4 and resource 5, as the other agents are covered by the remaining resources. If we would now change k to 2, we would need to add an additional grumpy agent. However, we cannot give a positive payoff to both this additional grumpy agent and all the regular agents at the same time. This is consistent with the fact that there is no solution for the k-SET COVER instance if $k = 2$.

2.7 Conclusion and Future Work

We have evaluated the effect of information disclosure and capacity optimization in a minority game designed to study crowding effects in public transport. The inclusion

Table 2.4: An example reduction from k-SET COVER to a model instance.

(a) An example instance of a k-SET COVER problem.

Set	e_1	e_2	e_3	e_4	e_5
A_1	×	×			
A_2				×	×
A_3		×	×		
A_4			×	×	
A_5	×				×

(b) The corresponding choice sets for $k = 3$

Agent	C_i
a_1	$\{\emptyset, \{1\}, \{5\}\}$
a_2	$\{\emptyset, \{1\}, \{3\}\}$
a_3	$\{\emptyset, \{3\}, \{4\}\}$
a_4	$\{\emptyset, \{2\}, \{4\}\}$
a_5	$\{\emptyset, \{2\}, \{5\}\}$
g_1	$\{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$
g_2	$\{\emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$

of heterogeneous agents and the availability of new information technologies such as smart phone apps and smart card ticketing systems pose many new challenges for the management of public transport systems. From the theoretical perspective there are questions to what extend observations for the original minority game, such as the relation between memory length and efficiency, still apply. From the practical perspective, the question is whether an operator can influence and manage the cooperation of the agents in order to stimulate the efficient utilization of the vehicles. We have conducted two simulation studies where we focused on the practical challenges.

In the first study we evaluate the effect of different information policies in a scenario where every agent uses at most a single trip every round. We find that disclosing more information attracts more agents, but that this comes at the cost of lower payoffs. This trade-off is influenced by the number of agents and the available capacity in the system. There is a clear impact of different agent types on the outcome of the simulations. A useful area for future research is to get a better understanding of the decision strategies employed by real travelers and how these compare to the simple decision strategies currently used in our model.

In the second simulation study, we evaluate the effect of adaptive capacity management in the context of railway transportation. Here the agents make a journey along a railway line. They have to choose a time to travel between an individually assigned origin and destination every round. As such journeys can cross multiple stops and thus overlap on the line, more complex patterns of agent interaction can emerge. We find that the number of rounds utilizations are recorded before capacities are re-optimized has an impact on the number of agents utilizing the system and their payoffs in the long run. Although the cost-benefit ratio for the operator was minimized when rescheduling every round, the cost per unit of payoff for all agents was greatest with rescheduling every 10 rounds. Further research is necessary to analyze the impact of the ratio between fixed costs and variable costs on these cost-benefit ratios.

Our studies show that we are able to evaluate and compare the effects of different models and policies for information and capacity. We observe that the amount of information available to public transport users, communicated by means such as smart phone apps, as well as the way historic information on vehicle utilization, such as can be collected by means of smart card technologies, can have a big impact on the number and satisfaction of public transport users. The question remains whether we can improve on the policies we evaluated. We think that policies that act on agents that repeatedly have a low payoff are an interesting area for further research.

To summarize, we have shown that our agent-based model is able to model complex interactions, in particular those related to crowding. This holds both for interactions among passengers within a public transport system and between the passengers and the public transport operator. We are able to observe an impact of the way passengers use information in their decision make and the amount of information provided by the operator, for example by using smart phone apps. The model can also be used to model and evaluate the impact of demand predictions based on smart card data for the purpose of capacity planning. With the inclusion of realistic behavioral models our model can be utilized to further study the impact of technology on the interactions within public transport systems.

3

Time Choice Data for Public Transport Optimization

Co-authors : Clint Pennings, Jan van Dalen and Leo Kroon
This is chapter is a working paper.

3.1 Introduction

As urbanization increases and working habits become more flexible, it is a reasonable fear that transportation systems will become more crowded and congested. Vehicle utilization is a complex issue to deal with for public transport operators. Having enough passengers in a vehicle makes a service profitable, while having too many passengers leads to decreased comfort and less attractive public transport. The resources necessary to cover the peak demand are typically not fully used outside the peak hours, leading to a low overall vehicle utilization and a relatively high cost per passenger mile.

With the advent of smart card ticketing technologies, public transport operators have obtained a large amount of microscopic data on passenger journeys (Pelletier et al., 2011). It is tempting to think that one simply has to analyze th data to learn everything there is to know about passenger behavior and to incorporate this knowledge into optimization models. Unfortunately, smart card data misses important aspects of passenger behavior, such as which alternative travel options have been considered by the passenger, the price sensitivity of the passenger and the potential to reschedule activities. As a result it is not known which decisions a passenger would have made if different alternatives had been available. However, it seems reasonable to assume one can overcome this issue if commuting passengers are the focus of the study, since those passengers generate enough data to make an educated guess about the choices they consider. Since commuting is the main cause for peak demands smart card data of commuters is a promising data source for the study of crowding. To the best of our knowledge it is an open question how to do this in a way that is useful for the operational planning models used by public transport operators.

In this chapter we discuss three major contributions to the study of crowded public transport systems in the big data era. First, we design an experiment that allows us to obtain data with a similar structure as smart card data, but with the added advantage that alternative choices, system parameters and the availability of information are fully controlled and can be manipulated for different groups of respondents. As we executed this experimental design with a large group of second year Bachelor students rather than a true random sample of the population, we refer to this study as a *quasi-experiment*. Second, we develop individual behavioral measures that can be utilized to analyze the behavior of passengers based on their choice data. Finally, we show that these measures can be successfully applied to create behavioral clusters which can be used to define behavior in agent-based models and segmentation of the customer base.

Crowded vehicles clearly cause discomfort for passengers, but it is not well understood at which point passengers change their travel behavior in response to crowded vehicles. Crowding has been considered by researchers from the field of transport policy (Li and Hensher, 2011), investigating the economic effects of crowding on passengers and how this should be incorporated in long term investments and network design. Crowding, however, is also a day-to-day matter in public transport operations. One way in which operators can deal with crowded situations is clever assignment of different vehicle types to optimize capacities (Fioole et al., 2006). Another way to deal with crowded situations is to adapt the frequency in which services are operated (Borndörfer et al., 2007).

In addition to scheduling based approaches, pricing incentives to seduce passengers to travel outside the peak hours have been considered (Link, 2004; Knockaert et al., 2012), although it was not received positively by public transport passengers. An even different angle is to provide information about the expected crowding level of each train service, e.g. via a smart-phone application, so passengers can re-plan their time of travel.

The difficulty with these approaches is that they will likely influence the behavior of passengers once implemented. Such a change in behavior will be observed by the operator, which may react by again adapting capacity while stochasticity in operations can cause fluctuations of the punctuality. This can cause new changes in passenger behavior and we obtain a feedback loop of passenger-operator interaction in the transportation system. Ideally, such interaction effects should be taken into account by the operator during service design. However, typical scheduling models assume demand to be fixed a priori and ignore interaction effects. One major reason for this is the fact that passenger behavior is typically very time consuming to measure and track, let alone include in the already challenging scheduling problems.

In earlier research presented in Chapter 2, we have developed a simulation framework that can be used to analyze the interactions between different choice strategies of passengers and capacity optimization strategies of the public transport operator. In the current research we have conducted a quasi-experiment in order to obtain a dataset that allows us to model iterative choice strategies based on longitudinal data, such as smart card data. Our quasi-experiment has a number of advantages over the direct use of smart card data. First, this way we have greater control over the experimental design, which allowed us to introduce a number of different experimental manipulations. Secondly, it is usually not possible for public transport companies to share actual smart card data with the research community because of privacy laws and competitive concerns. With the results from the quasi-experiments we can develop methodologies without these concerns and provide these methods to public transport companies at a later stage. As such we can easily share and discuss this dataset with other researchers. Furthermore, we also collected data on psychological traits, which can prove useful for more in-depth studies than are possible with only smart card data.

In this chapter, we explore a number of measures that allow us to analyze individual choice vectors. By using k-means clustering, we obtain four archetypes of passengers, the *stoic* who never switches, the *dualist* who switches between only two different alternatives, and finally the *moderate switcher* and *heavy switcher* which keep choosing from more than two travel options but with different frequencies. Based on these archetypes we can establish that the availability and quality of crowding information provided to passengers influences their behavior. We analyze how three different manipulations, the occurrence of disruptions, the accuracy of information and the crowding regime affect the occurrence of these different types of behavior. Based on this, we make recommendations for the development of optimization models for short term operational planning that takes the reaction of passengers in a crowded situation into account.

The remainder of this chapter is organized as follows: In Section 3.2 we discuss related literature. We introduce the design of the quasi-experiment in Section 3.3. In Section 3.4, we discuss the behavioral measures that are a the core of methodology we apply to analyze the data obtained by means of the quasi-experiment. In Section 3.5 we discuss the results of this analysis.

3.2 Related Work

Present study related to the analysis of smart card data in public transport serves as an important motivation for our methodology. The traditional approach to the modeling of transportation behavior involves estimating the economic value (typically expressed in a monetary value or a generalized travel time) of various sources of discomfort by means of stated choice experiments. While the utility functions obtained this way are extremely valuable, particularly in long term infrastructure

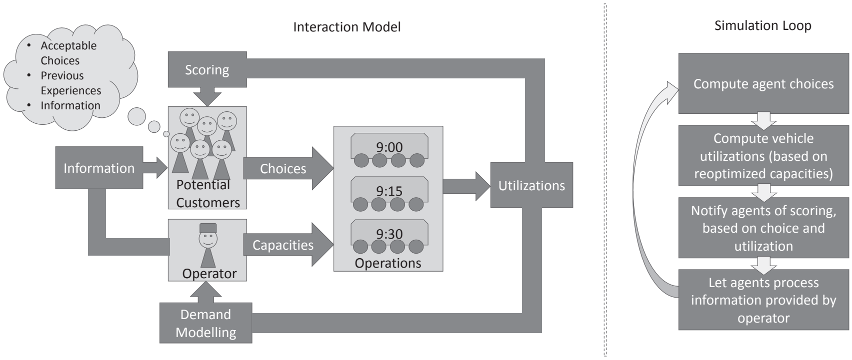


Figure 3.1: Overview of a simulation framework where the interactions between individual decision strategies and operator policies that influence demand can be simulated and analyzed.

planning and network design, they are specifically designed to model equilibria under the assumption that all participants in the transport system are able to find this equilibrium.

For operational planning on a day to day basis, it is questionable whether passengers are always able to find such an equilibrium, especially when small changes are made in the public transport system, when disruptions or delays occur, or when unusual high peaks of crowding occur.

In Chapter 2 we have developed a simulation framework where the interaction of passengers and public transport operators can be modeled based on arbitrary individual decision rules. This enabled us to analyze the behavior of simulated transportation systems on short to medium time scales (i.e. days to weeks) and other situations where rationality seems like an unrealistic assumption. We studied the impact of crowding information and the interaction between passengers who try to evade crowded situations and a public transport operator who is optimizing capacity based on crowding observations. A visual abstract of this research is presented in Figure 3.1. For the practical applicability of such models it is important to use behavioral rules for the agents that are validated by data from the passenger population. Since the current model only makes use of simple but unvalidated behavioral rules, a methodology that is able to use data from smart cards or similar data sources to calibrate agent decision behavior is a vital contribution to such models.

The optimization of the capacity offered by a public transport system also serves as an important motivation for our research. One important aspect of public transport supply is the frequency at which a certain service is operated. Methodologies for line planning are used to determine which routes to offer at which frequencies based on an origin destination matrix modeling the passenger demand (Borndörfer et al., 2007; Schöbel, 2012). Since line planning is typically done for long term horizons and it complicates the decision process of passengers, the approach was limited to *rolling stock allocation* (Fioole et al., 2006). This is a later step in the public transport planning process which has a direct impact on the service quality and crowding experienced by passengers. Nowadays, Netherlands Railways (NS) is able to produce a new rolling stock circulation model for the next day, which fits our goal to consider it for data driven short term public transport optimization.

We will now discuss these different fields in more detail.

3.2.1 Public Transport Policy Planning

The four step model (McNally, 2000) is the most common model for public transport demand modeling used in long term planning practice. It consists of four steps: Trip generation, trip distribution, mode choice and route assignment. The disadvantage of this model is that the planning region must be divided into fixed zones. Travel demand is then computed by estimating the number of trips between zones. A different approach is the use of agent-based modeling, where transport demand is modeled using individual entities that can have more intricate travel plans and decision strategies. Although the four step model is most commonly used by practitioners, a number of large scale agent-based models have been successfully developed for the purpose of assisting (public) transport authorities in their policy development (Balmer et al., 2006).

An important aspect of models used for long term policy and decision making is the use of *utility functions* which are used to predict how satisfied travelers will be with certain journeys, modes and prices. A widely used survey technique for this is discrete choice modeling (Bierlaire, 1998), where respondents are shown *different* alternative travel options and are asked to pick the one they prefer from certain carefully selected combinations. Based on the reported preferences, a utility function is fitted on the acquired data. This survey technique differs from our quasi-experimental design in that we present the same choices throughout the experiment, but vary the outcome presented *after* the choice has been made. This makes our approach less suited for the estimation of utility functions, but enables us to study choice behavior over time.

An important model closely related to game theory for modeling how individuals will travel trough a transportation network for a given origin destination matrix and flow-dependent arc costs is the Wardrop equilibrium (Wardrop, 1952). It is computed for a network with flow-dependent travel times where rerouting (part of) the flow for a single origin destination pair will never lead to a faster route for that pair. The last step of the four step model usually generates a variation of this equilibrium, where travel times are replaced by general utilities.

3.2.2 Smart Card Data Analysis

Research based on data collected via smart card ticketing technologies has gained a lot of popularity since many public transport systems have adopted these systems

(Pelletier et al., 2011). One of the advantages of smart card data is that repeated use of the same smart card can link multiple journeys together. Depending on the precise implementation of smart card technologies, it can be challenging to link smart card transactions to specific public transport services in time and space. Some systems only have a check-in at a station and as a result it is necessary to deduce likely check-out stations (Trépanier et al., 2007). In systems where both check-in and check-outs occur, data from conductors checking smart card can be used to validate different methods for linking smart card transactions to public transport services (van der Hurk et al., 2015). Other researchers have worked on the analysis of mobile phone location data, but such data is different in its temporal resolution and spatial resolution and as a result we cannot merely substitute mobile phone data with smart card data. For a discussion about travel behavior analysis based on mobile phone data, we refer the reader to Chen et al. (2016).

3.2.3 Complex Systems Science

Many fields of research are associated with the label *complex systems science*, such as chaos theory such as the Lorenz-system (Lorenz, 1963), cellular automata such as Conway's Game of Life (Gardner, 1970) and processes that grow large networks with certain properties, such as the small-world and scale free networks (Watts and Strogatz, 1998; Barabási and Albert, 1999) as well as agent-based modeling.

A number of models of *complex systems science* have contributed to the field of transportation. The Nagel-Schreckenberg (Nagel and Schreckenberg, 1992) model is a cellular automaton in which cars behave according to two simple driving rules. A car speeds up to a certain speed if no other car is in front of it and brakes when the distance to the next car gets too small. With these rules, a group of cars interacting on a highway may end up in a traffic jam under certain conditions.

An model which deals with coordination and utilization of individual agents is the El-Farol bar game (Arthur, 1994). In this model a population of agents has to decide every week whether to go to a bar or stay at home. Going to the bar yields a positive or negative payoff depending on whether the number of attendees is beyond a certain crowding threshold. Staying at home yields a neutral payoff. In the original paper it was observed that although the aggregate behavior of the population converges, individuals keep switching their strategies. This complex system has also been studied by the name *Minority Games* (Challet et al., 2000). The *minority game* perspective has been applied to route choice in car traffic (Bazzan et al., 2000; Wahle et al., 2000). Selten et al. (2007) discuss an experiment where respondents had to select one of two routes for multiple rounds. Their approach puts the respondents in a computer lab and reported the outcome of a round based on the responses of all respondents. The drawback of this approach is that it requires all respondents to be present and wait for each other, limiting the scale at which data can be collected. The quasi-experiment discussed in this chapter overcomes this issue by developing an interactive experiment that can be performed by a single respondent at a time. To the best of our knowledge we are the first to consider minority games in the context of public transport, which has different characteristics when compared to car traffic. Our simulation study in Chapter 2 combines the ideas from minority games with a public transport operator which re-optimizes capacities based on the observations in the minority game.

3.3 Design of the Quasi-Experiment

In order to gain insight into the choice behavior of passengers in an iterated setting where the same travel decision has to be made many times, we have developed an interactive quasi-experiment modeling a daily commuting scenario. This quasi-experiment consists of three main phases: (1) Travel decisions without crowding indicators for 20 rounds. (2) Travel decisions with crowding indicators for 20 further rounds. (3) Survey questions on personality traits and what the respondent focused on during the experiment, as well as personal experience with public transport.

We conducted the experiment among more than 500 second year bachelor students, a relevant group in the Netherlands as they make intensive use of public transport and are responsible for a large part of public transport demand. The quasi-experiment was held using an online environment. In this environment the respondents were repeatedly asked to choose a travel mode (train or car), and in case of the train their preferred time of travel. The verbatim text of the experiment is available in Appendix B.1.

Respondents were incentivized by the use of their collected data in an important assignment within a mandatory bachelor course on statistics. Furthermore, respondents could partake in a quiz where they could win a gift coupon. The outcome of the quiz was not related to the answers given in the quasi-experiment.

The travel time decisions during the first two phases of the experiment were based on a realistic commuting scenario between the Dutch cities of Utrecht and Den Bosch, one of the most frequented commuting trips in the railway network of the Netherlands. Between these cities train travelers can choose between either one of the local trains, which stop at every station in between, or one of the fast trains that only stop in the big cities. Both stations are situated in the city centers. Traveling between the city centers by car involves leaving and entering the city center and driving along the highway for the remaining part. As a result, the fast train can be a faster option if both home and work locations are situated within the city centers. This is also the scenario communicated to the respondents. The travel choices offered to the respondents are presented in Table 3.1.

After making a choice, the respondent receives information on how the journey turned out and how crowded the train was. The possible responses for crowding are presented in Table 3.2. An example response is:

Table 3.1: The 8 possible travel options. The respondents were informed that they should be at work 8:30 the latest, with their home and work locations at negligible distance from the stations.

Choice	Mode	Departure Time	Duration (minutes)	Arrival Time
1	Train	7:11	46	7:57
2	Train	7:23	30	7:53
3	Train	7:38	29	8:07
4	Train	7:41	46	8:27
5	Train	7:53	30	8:23
6	Train	8:08	29	8:37
7	Car	7:30	55	8:25

Table 3.2: The three possible types of feedback presented after a travel option has been picked

0	You sat comfortably and were able to work in the train.
1	You were able to work in the train for most of the journey.
2	You had to stand and were unable to work.

Your scheduled arrival time was 7:53. Your actual arrival was 7:56. You were able to work in the train for most of the journey.

Respondents are then asked to indicate on a scale from 1 to 5 how satisfied they are with the outcome of their choice. These two steps, making a travel choice and reporting satisfaction based on the provided feedback, are repeated for each round. During the first and second phase of the experiment the respondent fills in twenty rounds, so forty in total.

During the second phase of the experiment the respondent has to make the same decision again for twenty rounds, but this time a crowding indicator (see Figure 3.2) is presented, predicting the crowding of the the train options.

We present a formal definition of the data used during the experimentation as well as the implementation details of our experimental manipulations in Section 3.3.2, but first we motivate and discuss the different experimental manipulations separately.

3.3.1 Experimental manipulations

In order to compare the choices of different respondents, we presented the same results (arrival time and crowding) to the respondents who make the same choices, if they are within the same experimental manipulation group. This also has the

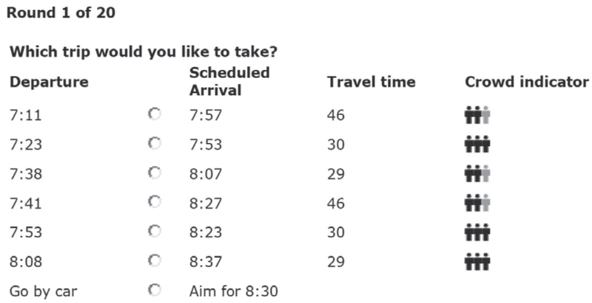


Figure 3.2: Presentation of the different choices with crowding indicators in the quasi-experiment

advantage that an individual respondent does not have to wait for the travel choices of the other respondents during the experiment. Typically, experiments based on direct interactions between the respondents, e.g. Selten et al. (2007), are much more costly to set up (i.e. they require a computer lab) and allow only for a small number of participants.

With a large number of respondents we had the opportunity to introduce three experimental manipulations. Each manipulation was applied to a respondent with probability 0.5. As a result we are able to compare the respondents who did not receive a certain manipulation with the ones which did receive that manipulation. As there are three manipulations which a respondent can have or not, there are eight different combinations of manipulation assignments which we will refer to as *groups*. The three different manipulations are the following: One group only observes small delays to the planned travel time, while another group might run into a disruption of half an hour in a crowded vehicle (Occurrence of Disruptions). During the second phase of the experiment one group will see crowding indicators that perfectly predict what will happen in the train, while the other group gets crowding indicators that are random (Quality of Information). One group will be discouraged to keep using the same travel option due to increased crowding of the last choice, while the other group has to deal with purely random crowding (Reactive Crowding). One group will be discouraged to keep using the same travel option due to increased crowding of the last choice while the other group has to deal with purely random crowding (Reactive Crowding). This manipulation can be interpreted as a public transport operator which moves capacity away from the travel option last chosen by the respondent to the adjacent travel options and are thus capacity decisions that are bad for a respondent who wants to stick to a single choice.

As soon as a respondent entered our system they were randomly allocated to one of these manipulations. This was not communicated to the respondent in any way, so it is reasonable to assume that differences in behavior between groups are caused by the manipulations. The motivation to include the occurrence of disruptions is that disruptions often cause very crowded situations in public transport, so this is important to consider when studying crowding behavior. The quality of information is something which operators have to predict, so in the real world there is always an error margin in these predictions. Including it as a manipulation allows us to compare the effect of errors in the crowding indicators on choice behavior against a best case scenario. The motivation to include reactive crowding as a manipulation is that it allows us to analyze whether increased crowding levels are give a greater incentive to change travel times than purely random crowding levels.

3.3.2 Formal definition of the interactive data and procedure of the quasi-experiment

In order to formally define the interactive nature of the experiment, we need to define both the data presented to the respondents and the computations we used to analyze the experiment. We begin by defining a vector of the arrival times based on Table 3.1 in the following way:

$$\text{arr} := [7{:}57, 7{:}53, 8{:}07, 8{:}27, 8{:}23, 8{:}37, 8{:}25]$$

During the experiment, the respondent gives feedback during twenty rounds. The feedback given to the respondent in each of these rounds was generated using a number of matrices. These matrices are drawn a single time prior to the quasi-experiment. As a result, the feedback and crowding indicators observed by the respondents are based on the same data matrices. These data matrices are defined as follows:

The *crowding matrix* $U \in \{0, 1, 2\}^{6\times 20}$ for train options. An entry $U_{c,r}$ is the crowding during round r for choice c and either 0 (low crowding), 1 (moderate crowding) or 2 (high crowding) where 0 is low crowding and 2 is high crowding. The crowding was drawn uniformly and independently at random. Note that the crowding matrix contains no crowding for the car choice, $c = 7$.

The *delay matrix* $D \in \mathbb{Z}^{7\times 20}$. An entry of the delay matrix $D_{c,r}$ contains the delay of choice c during round r, which comes on top of the travel time communicated to the respondents. Each entry was computed by taking the absolute value of an independent draw from the normal distribution $|\mathcal{N}(0, 3)|$ for the train choices. In the Dutch network, more than 90% of the trains arrive "on time", which means they are delayed by at most five minutes. For the car choices the delay matrix is used to model stochasticity due to traffic lights and minor traffic. The distribution $|\mathcal{N}(0, 3)|$ was used in case the disruption matrix DI (which we define next) contains a 1 for the car option and $|\mathcal{N}(0, 6)|$ was used otherwise. We refer to the delay matrix that used $|\mathcal{N}(0, 3)|$ by D for the car choices, while we refer to the $|\mathcal{N}(0, 6)|$ matrix by D'. Thus, we assume that the standard deviation of delays of the train is three minutes and of the car is either three or six minutes.

The *disruption matrix* $\mathrm{DI} \in \{0,1\}^{7\times 20}$. An entry of the disruption matrix $\mathrm{DI}_{c,r}$ contains 1 if a disruption occurred for choice c during round r. If a respondent faces a disruption this implies an increase of 30 minutes in the travel time. The entries were drawn independently from the Bernoulli distribution $\mathcal{B}(0.1)$ for all the train choices and from $\mathcal{B}(0.3)$ in case of a car choice, were a disruption represents an accident or a severe traffic jam. The probability was chosen in such a way that in expectation a respondent runs into a disruption twice when travelling by public transport. The probability for disruptions in car based transport are higher as we model route with very high demands which often result in traffic jams, while rail punctuality is typically quite high in the Netherlands.

The *random information matrix* I, which is only relevant during the second phase. This matrix is generated in the same way as the matrix U. This way, the crowding indicator levels are generated according to the same distribution as the real crowding. In case we want to provide accurate information to a respondent, the crowding matrix U is used and in the base case there is a $\frac{1}{3}$ probability that an entry in U is accurate and equal to I and a $\frac{2}{3}$ probability that it is not.

During the second phase of twenty rounds, respondents were presented with crowding indicators before making a choice, representing the predicted crowding of each train choice. The crowding indicator indicated the expected crowding level consisting of either one, two or three symbols representing a person.

Our three experimental manipulations yield $2^3 = 8$ respondent groups, denoted by a triplet (δ_j, q_j, ρ_j). The three manipulations are referred to as occurrence of disruptions, quality of information and reactive crowding.

The first manipulation, occurrence of disruptions, consists of the occurrence of large disruptions, indicated by $\delta_j \in \{0,1\}$ for respondent j. One group of respondents, $\delta_j = 0$, were incidentally confronted with a large disruption and the other group only faced small delays.

The second manipulation, quality of information, consists of the quality of information during the second phase of the experiment, indicated by $q_j \in \{0,1\}$ for respondent j. One group, $q_j = 0$, received accurate crowding level information and the other group received random information.

The third manipulation, reactive crowding, consists of the relation between the crowding level and the prior choice, indicated by $\rho_j \in \{0,1\}$. One group, $\rho_j = 0$, experienced a purely random crowding level and for the other group the crowding level was partly dependent on the previous choice.

Reactive crowding was implemented as follows. If a respondent picked a certain train journey the previous round, the "*crowd*" will be attracted to this option, making that option more crowded during the current round. This increase in crowding is modeled to be due to a shift of the crowd from travel options close in time. As such, the crowding of the adjacent (with respect to the choice number) travel options is decreased by one. As a result, a respondent who only cares about avoiding the crowding and who is in the $\rho = 1$ group has an incentive to keep switching from its current travel option. In case the choice during the previous day was either the car or does not exist, the "*regular*" crowding level, as defined in the matrix U, is used. Given a base crowding level c described in U, the function t_j transforms this crowding level c to the manipulated crowding level based on the current chosen travel option o and the previously chosen option o' when $\rho_j = 1$ and is the identity function in case $\rho_j = 0$. In case the crowding level is adjusted, we make sure it will be one of the three crowding levels by means of the min and max functions. It is formally defined as follows:

$$t_j(c, o, o') := \begin{cases} c & \text{if } o' = 7 \\ \min\{2, c+1\} & \text{if } \rho_j = 1 \wedge o = o' \\ \max\{0, c-1\} & \text{if } \rho_j = 1 \wedge |o - o'| = 1 \\ c & \text{otherwise} \end{cases} \tag{3.1}$$

This function reflects the idea that the crowding level is just a random draw in case the respondent is in a $\rho_j = 0$ group, or has selected the car during the previous round. If the respondent is in the $\rho_j = 1$ group, the crowding level is modified depending on the choice of the previous round.

Example 3.1. *Suppose that during round* 5 *the respondent* j *picked choice* 4*. During round* 6*, by chance the crowding of travel options* 3, 4 *and* 5 *is defined to be* 1*. If* $\rho_j = 0$*, i.e. respondent* j *has random crowding, the reported crowding will be* 1 *for all these choices. However, if* $\rho_j = 1$*, i.e. respondent* j *has reactive crowding, the reported crowding for options* 3 *and* 5 *will be* 0*, while it will be* 2 *for travel option* 4*.*

While one can argue that this crowding behavior may not be realistic in actual transport systems, this rule is designed to have two properties which we believe to be important for the experiment: 1) the rule should be deterministic in a sense that two respondents who make the same choices observe the same crowding levels, as this is important for a fair comparison between correspondents; and 2) the crowding should be noticeable by the respondents, as one of the important questions is whether crowding interactions affect the choice behavior of the respondent *at all*. Even if our crowding behavior would be exaggerated, it is still helpful to determine whether or not there is an effect.

When the quasi-experiment is currently in the second phase, the respondent is shown a crowding indicator before making a choice, as is visualized in Figure 3.2. Thus we need to compute the value of the crowding indicator for each train choice during every round r before presenting the possible travel choices to respondent j. We compute the function $ind_j(r, o, o')$ for each chosen travel option o based on the travel option chosen by respondent j during the previous round, denoted by o'. This computation depends on j's quality of information q_j and is defined as follows.

$$ind_j(r, o, o') = \begin{cases} t_j(U_{r,o}, o, o') & \text{if } q_j = 0 \\ t_j(I_{r,o}, o, o') & \text{otherwise} \end{cases} \tag{3.2}$$

The sole purpose of this function is to switch between the crowding levels, and the random draws that are considered to be the predicted crowding levels, based on the q_j property of respondent j.

For both the first and second phase, the respondent is provided with feedback after a travel option o has been picked. First, we will calculate the amount of travel delay $trd_j(r, o)$ during round r for respondent j, which is added to the regular travel time in the feedback. This computation depends on whether respondent j experiences any disruptions, denoted by δ_j.

$$trd_j(r, c) = \begin{cases} D_{r,c} + 30(1 - \delta_j)\, DI_{r,c} & \text{if } c < 7 \vee \delta_j = 1 \\ D'_{r,c} + 30\, DI_{r,c} & \text{otherwise} \end{cases} \tag{3.3}$$

Finally, the crowding level $cr_j(r, c, c')$ experienced by respondent j during the chosen travel option o and the previous travel option o' is computed. The feedback provided to the respondent between picking a travel option and recording the satisfaction contains either one of three remarks from Table 3.2. In order to obtain this level, the crowding transformation function t_j is applied to the current utilization $U_{r,c}$, except in the case of a heavy disruption. If there is a disruption, we assume that capacity is insufficient and the train will be at the maximum crowding level. The formal definition of cr_j is as follows:

$$cr_j(r, c, c') = \begin{cases} 2 & \text{if } trd_j(r, c) \geqslant 30 \\ t_j(U_{r,c}, c, c') & \text{otherwise} \end{cases} \tag{3.4}$$

Based on these formulas and matrices, we formally specify the loop of the quasi-experiment for the two interactive phases of the experiment, as is presented in Algorithm 1.

3.4 Methodology and Analytics

After having obtained the data, we analyze it with a methodology that in principle can also be applied to data collected by operators. We will apply the following three steps:

1. Analyze descriptive statistics of the entire population, in particular the number of times the different travel choices have been selected and the number of respondents who change their travel choice compared to the previous round.

2. Derive and analyze behavioral measures for individual respondents based on the selected choices and the experimental data linked to these choices.

3. Obtain behavioral profiles by clustering the behavioral measures of the previous step.

Algorithm 1: Loop for the two interactive phases of the quasi-experiment

```
Data: Loop for respondent j during either the first or second phase
r ← 1 ;
c′ ← −1 ;
while r ⩽ 20 do
    Print Table 3.1 ;
    if Phase = 2 then
        for k := 1 to 6 do
            Show indicator with ind_j(r, k, c′) symbols for choice k.
    c ← Chosen User Input ;
    Print "Your scheduled arrival time was " + arr_c ;
    a ← arr_c + trd_j(r, c) ;
    Print "Your actual arrival was " + a ;
    u ← cr_j(r, c, c′) ;
    if a ⩾ 30 then
        Print "You experienced a disruption."
    if u = 0 then
        Print "You sat comfortably and were able to work in the train."
    else if u = 1 then
        Print "You were able to work in the train for most of the journey."
    else
        Print "You had to stand and were unable to work."
    r ← r + 1 ;
    c′ ← c ;
```

After collecting the answers of the respondents, we plotted the choices, switching behavior and reported satisfaction to visualize the results over the course of the experiment. This yields a high level description of the collected data.

In order to distinguish different types of behavior, we first need to define proper behavioral measures. We focus on measures that are based on data already available at many public transport operators, specifically the travel choice picked by a passenger, delay information, the utilization of vehicles and possibly the satisfaction reported by the respondents. Of these, microscopic satisfaction scores are the hardest to obtain for the operator. However, recent versions of the smart phone app of Netherlands Railways (NS) allows passengers to report how satisfied they were with their current journey.

We will first define a number of behavioral measures in Section 3.4.1. We perform two types of analysis using these measures. First we compare the behavior of the different experimental manipulation groups among the different phases. Second, we apply a clustering method in order to discover different behavioral profiles within the data. We then compare how these behavioral profiles are distributed among the different experimental manipulation groups and in the first and second phase of the experiment.

In order to isolate the effects of different experimental manipulations and among phases, we simply split the data according to a manipulation, perform the measures and use the standard t-test to compare whether the two groups are significantly different. Since we perform 9×9 of such comparisons, we choose 0.01 as a threshold for significance and report the p-values to know in which cases we are close to this bound.

In order to distinguish different types of behavior, we propose to use k-means clustering. The question with k-means is which number of clusters to choose. We prefer to solve this with a scree-plot (explained in Appendix A), aiming to keep it reasonably small for easier interpretation. The scree-plot displays the within-cluster sum of squares against the number of clusters. The plot is visually inspected to find a sharp bend, after which the number of clusters is chosen to be just before the point after which adding more clusters has a relatively small impact on the sum of squares. We consider the behavior during the first and second phase separately for the full population.

After assigning a behavioral type to each respondent for the first phase and the second phase, we can analyze whether the number of times each behavioral type occurs differs between the experimental groups when compared to the full population. We use an exact binomial test to compute how likely the observed distributions could have been generated at random as a statistical test.

Finally, we analyze the change in behavioral types from the first to the second phase, by regarding the fractions of respondents who transfer from first phase clusters to second phase clusters.

3.4.1 Behavioral Measures

In this section, we define a number of behavioral measures that can be computed for a single respondent. We had the following goals and preferences in mind while defining these measures:

First, preferably the measures should be based on the choices made by an individual respondent, since we are looking for a way to describe the behavior of individuals.

Second, the measures should be based on information available or reasonably obtainable by a public transport operator, such as information on delays and disruptions (i.e. the realized timetable), the utilization rate of the vehicles (by combining smart card data and a rolling stock schedule) or the information communicated to passengers via apps or websites.

Third, the measures should be as simple as possible. Although advanced time series techniques could be applied, we prefer measures that look back a single round for reasons of simplicity.

Fourth and finally, the measures should have a scale that is easy to interpret, preferably $[0, 1]$.

In order to introduce the measures we picked based on these preferences, we will use the notation introduced in Section 3.3.2. Since we want to measure the behavior of a single respondent, we first assume that the respondent number j is fixed, as are the choice vector C and the satisfaction vector S. The measures are defined on either one or both of these vectors and a window k. The window restricts the number of observations that are used, so if we want to compute a certain measure for only the first ten rounds, we use a k of 10. Furthermore, while some measures such as the number of switches can be computed without specific knowledge of the respondent group, others require the specific group of j, i.e. q_j, δ_j and ρ_j. Only the measures that require knowledge of the respondent group will be parametrized by j.

The choices of a certain respondent j that were recorded by the online environment are stored in a vector $C = [c_1, c_2, \ldots, c_{20}]$ and the satisfaction scores in a vector $S = [s_1, s_2, \ldots, s_{20}]$. Formally, the set $\mathcal{C}$ of all possibly choice vectors is defined as $\mathcal{C} = \{1, 2, \ldots, 7\}^{20}$. In addition, a vector $X = [x_1, x_2, \ldots x_{19}]$ indicating whether the next choice will be different from the current one. Formally

$$x_r = \begin{cases} 1 & \text{if } c_r \neq c_{r+1} \\ 0 & \text{if } c_r = c_{r+1} \end{cases}$$

Note that X solely depends on C and is only defined in order to simplify the mathematical expression of some of our measures.

In order to keep the measures as simple as possible, we will mostly focus on the switching behavior of a respondent.

A central measure for the behavior of a respondent is the probability the respondent switches. This can be computed by dividing the number of switches by the number of rounds. We will call this measure sp. This measure is formally defined as follows:

Table 3.3: Overview of behavioral measures used for analysis

Measure and short name	Description of measure
Crowd avoidance: ca	The probability of switching in case of high crowding or to remain in case of low crowding during the previous round. Formally, it can be viewed as $\Pr((x_i = 1 \mid \mathrm{cr}_j(i, C) = 2) \cup (x_i = 0 \mid \mathrm{cr}_j(i, C) = 0))$
Reactive switching coefficient: rsc	The probability of switching probability when the satisfaction was either 4 or 5 or to remain if the satisfaction was either 1 or 2 during the previous round. Formally it can be written as $\Pr((x_i = 1 \mid s_i < 3) \cup (x_i = 0 \mid s_i > 3))$.
Adjustive reactive switching coefficient: rsca	The probability of switching if the satisfaction during the previous round was higher than the average observed satisfaction of respondent j. Its formal definition is the same as the definition of rsc, where the constant 3 is replaced with the observed average.
Delay responsive switching: dr	The probability of switching when the delay of the previous round was higher than the average witnessed delay or to remain at the previous travel option otherwise. Formally it can be regarded as: $\Pr((x_i = 1 \mid \mathrm{trd}_j(i, c_i) \geqslant \mathrm{avg}_{\mathrm{trd}_j}(i)) \cup (x_i = 0 \mid \mathrm{trd}_j(i, c_i) < \mathrm{avg}_{\mathrm{trd}_j}(i)))$
Last switch: ls	The index of the round the last switch occurred divided by the total number of rounds considered. Indicates the time a respondent sticks to a single choice.
Minimum time for a choice set cardinality of 2: mtsc_2	The index of the last round after which only two choices are considered. Indicates the time someone restricts their choices to two options.
Average satisfaction: avg_S	The average of the satisfaction scores reported by the respondent.
Maximum streak: maxstr	The longest sequence of rounds where no switch was observed. Is an indication of the sequential stability of a respondents choices.
Sensitivity to information: si	Measures whether a respondent seems to ignore information completely or whether a respondent always choosen for an option with a certain type of information. Can only be computed in the second phase of the experiment.

$$\mathrm{sp}(C, k) := \frac{1}{k} \sum_{r=1}^{k} x_r$$

A primary design principle behind many of our proposed behavioral measures is that we want to capture whether a certain event during one round of the experiment will have an influence on the switching behavior of the respondent for its choice during the next round. Let us consider the probability that an individual will switch in round i given a choice set, which can be written as

$$\Pr(x_i = 1) = \mathrm{sp}(C, i)$$

Now suppose that event E occurred during round i. A typical way to detect whether this event affects the switching behavior is to check whether switching is independent from the event E, i.e. if

$$\Pr(x_i = 1|E) \neq \Pr(x_i = 1)$$

In case of inequality, it is also useful to know whether $\Pr(x_i = 1|E)$ is greater than or equal to $\Pr(x_i = 1)$.

Through application of Bayes' theorem, we can derive

$$\Pr(x_i = 1|E) = \frac{\Pr(E|x_i = 1)\Pr(x_i = 1)}{\Pr(E)} = \frac{\Pr(E|x_i = 1)}{\Pr(E)} \mathrm{sp}(C, i)$$

By application of Bayes' theorem we can see that when $\Pr(x_i = 1|E) > \Pr(x_i = 1)$ holds, we should also observe that the fraction of $\Pr(E|x_i = 1)$ divided by $\Pr(E)$ is greater than 1. When $\Pr(x_i = 1|E)$ is smaller than $\Pr(x_i = 1)$, the fraction should be strictly smaller than 1.

Thus, this fraction gives useful information about the influence of an event on the switching probability. However, the scale of this value is $[0, \mathrm{sp}^{-1}]$ and thus dependent on sp. As such it is not suitable to compare different respondents, since the base rate sp will likely be different for these respondents. For that reason, we prefer to use measures based on $\Pr(x_i = 1|E)$ instead. We developed the following measures that fit this general patterns, which are presented in the first rows of Table 3.3.

In addition to the effect of a certain event on the switching probability, we also defined a number of measures on other aspects of behavior. The following measures all have a natural maximum (e.g. the number of rounds considered) which can be used to normalize them to the $[0, 1]$ scale. These measures are presented in the bottoms rows of Table 3.3.

The formal definitions of these measures are discussed in Section 3.4.2.

3.4.2 Formal Definitions of Behavioral Measures

We will introduce some definitions in order to simplify the expressions of our behavioral models. We define versions of the cr_j and ind_j that take the choice vector C as input instead of two separate choices:

$$\mathrm{cr}_j(r, C) = \begin{cases} \mathrm{cr}_j(r, c_r, -1) & \text{if } r < 2 \\ \mathrm{cr}_j(r, c_r, c_{r-1}) & \text{otherwise} \end{cases}$$
$$\mathrm{ind}_j(r, C) = \begin{cases} \mathrm{ind}_j(r, c_r, -1) & \text{if } r < 2 \\ \mathrm{ind}_j(r, c_r, c_{r-1}) & \text{otherwise} \end{cases}$$

Sometimes it is necessary to ignore the data of rounds where the respondent chooses to go by car, since there is no information or crowding in case of car use. For this purpose we define a function that computes an index set of the rounds where the respondent chooses to go by public transport as follows:

$$\mathrm{PT}(C, k) := \{r : 1 \leqslant r \leqslant k, c_r \neq 7\}$$

In order to count how often the respondent traveled by public transport, we use the notation $\#\,\mathrm{PT}(C, k)$ to refer to the cardinality of the set.

We will now introduce a type of notation that is not common: The bracketed notation. This notation is useful to count how often a certain logical condition holds without resorting to specific set building notation. An expression $[a = b]$ has a value of 1 if the expression (in this case $a = b$) within the brackets is true, while it has value 0 otherwise.

In order to understand how crowding affects the switching behavior of an individual, we consider how the prior probability of switching, $\Pr(x_r = 1)$ will be affected by a crowded public transport experience. In order to do so, we can compare the conditional probability of switching *in case* the crowding was high, to the prior crowding probability. If the conditional probability is higher, the individual has a tendency to switch in case of crowded situations. As such, we are interested in the situation:

$$\Pr(x_r = 1 | \, \mathrm{cr}_j(r, C) = 2) > \Pr(x_r = 1)$$

If we use $\mathrm{sp}(C, k)$ as an estimate for $\Pr(x_r = 1)$, we can apply the theorem of Bayes to rewrite this to:

$$\frac{\Pr(\mathrm{cr}_j(r, C) = 2 | x_r = 1)}{\Pr(\mathrm{cr}_j(r, C) = 2)} > 1$$

Let us formally define this fraction as $\mathrm{ca}_j^{\mathrm{high}}$:

$$\mathrm{ca}_j^{\mathrm{high}}(C,k) := \frac{\sum\limits_{r \in PT(C,k)}^{k} [x_r = 1][\mathrm{cr}_j(r,C) = 2]}{\sum\limits_{r \in PT(C,k)} [x_r = 1] \cdot \sum\limits_{r \in PT(C,k)} [\mathrm{cr}_j(r,C) = 2]}$$

Similar, we expect that if someone is trying to avoid crowds, a low crowding during the current round will decrease the probability of switching. We are thus also interested in whether the following holds:

$$\Pr(x_r = 1 | \mathrm{cr}_j(r,C) = 0) < \Pr(x_r = 1)$$

We can apply Bayes' theorem in a similar way to obtain a fraction that should be smaller than 1 in case of crowd avoidance. We will define this ratio as $\mathrm{ca}_j^{\mathrm{low}}$ and compute it as follows:

$$\mathrm{ca}_j^{\mathrm{low}}(C,k) := \frac{\sum\limits_{r \in PT(C,k)}^{k} [x_r = 1][\mathrm{cr}_j(r,C) = 0]}{\sum\limits_{r \in PT(C,k)} [x_r = 1] \cdot \sum\limits_{r \in PT(C,k)} [\mathrm{cr}_j(r,C) = 0]}$$

Both $\mathrm{ca}_j^{\mathrm{high}}$ and $\mathrm{ca}_j^{\mathrm{low}}$ are very useful to update the prior switching probability, given the witnessed crowding level, but they have the disadvantage of having a scale which depends on $\mathrm{sp}(C,k)$. In order to compare and cluster different individuals, it is better to have a measure on a $[0,1]$ scale, which has a value of 1 in case $\mathrm{ca}_j^{\mathrm{high}}(C,k)$ is $\mathrm{sp}(C,k)^{-1}$ and $\mathrm{ca}_j^{\mathrm{low}}(C,k)$ is 0, and has a value of 0 if it is the other way around. In order to achieve this, we propose a single behavioral measure for *crowd avoidance*, defined as follows:

$$\mathrm{ca}_j(C,k) := \frac{\sum\limits_{r \in PT(C,k)} x_r[\mathrm{cr}_j(r,C) = 2] + \sum\limits_{r \in PT(C,k)} (1 - x_r)[\mathrm{cr}_j(r,C) = 0]}{\sum\limits_{r \in PT(S,k)} [\mathrm{cr}_j(r,C) \in \{0,2\}]}$$

Another indication of switching behavior, is the latest round ls during which a switch occurred. The lower this value is, the more likely the respondent has decided upon a fixed behavioral pattern, making the same choice every round. One way in which this measure can be defined is as follows:

$$\mathrm{ls}(C,k) := \underset{1 \leqslant r \leqslant k}{\mathrm{argmax}} \{r x_r\}$$

We can generalize this concept by analyzing the size of the choice set the respondent still considers. If we assume that a strategy of the respondent is to try many different options during the beginning of the experiment and then gradually restricts the relevant choices to a smaller set, it is useful to know from which round the respondent is only using only l different choices. We call this measure the *minimum time for a choice set cardinality*, and is parameterized by an additional parameter l, which specifies the maximum number of choices we allow the respondent to use before considering the restriction of the choice set small enough. It can be computed as follows:

$$\mathrm{mtsc}_l(C,k) := \mathrm{argmin}_{1\leqslant r\leqslant k}\{[\#\{c_i : r \leqslant i \leqslant k\} \leqslant l]r\}$$

Yet another type of switching behavior depend on the relation between reported satisfaction and the decision to switch. For this, we use the fact that the satisfaction is reported on a five step scale, so $S_i \in \{1, 2, \ldots, 5\}$ for any i. The tendency of a respondent to switch after a low satisfaction score and to stay with the same choice after a high satisfaction the *reactive switching coefficient* rsc:

$$\mathrm{rsc}(C,k) := \frac{\sum_{r=1}^{k} x_r[s_r < 3] + \sum_{r=1}^{k}(1 - c_r)[s_r > 3]}{\sum_{r=1}^{k}[s_r \neq 3]}$$

A potential issue with this measure is that a certain respondent may not use the whole range of possible satisfactions. In such a situation, the notion of a high or low satisfaction used within the rsc measure may not be appropriate for this specific correspondent. To overcome this, we can use the average satisfaction reported by the respondent as a threshold, instead of just taking 1 and 2 as a low satisfaction, and 4 and 5 as high satisfactions. The drawback of this is that we assume we know this average beforehand, something which may not hold in a setting where measurements are performed continuously. We obtain this different measure, the *adjusted reactive switching coefficient* rsca as follows:

$$\mathrm{rsca}(C,k) := \frac{1}{k}\left(\sum_{r=1}^{k} x_r[s_r \leqslant \mathrm{avg}_S(r)] + \sum_{r=1}^{k}(1 - x_r)\left[s_r > \mathrm{avg}_S(r)\right]\right)$$

where $\mathrm{avg}_S(k)$ just computes the average of the first k values of S:

$$\mathrm{avg}_S(k) := \frac{1}{k}\sum_{r=1}^{k} S_r$$

The *sensitivity to information* si_j is an aspect of behavior which is relevant during the second phase of the experiment, where we measure the extent in which information affects the behavior of a respondent. In order to come up with a definition for si_j we will first introduce a function that simply counts how often the respondent chooses a travel option for which a specific crowding indicator was shown. This function is called the *indicator count* and is defined as follows:

$$\mathrm{ic}_j(C, k, i) := \sum_{r \in PT(S,k)} [\mathrm{ind}_j(r, C) = i]$$

If we assume that a respondent is completely insensitive to information, we can hypothesize that the respondent chooses as if no information is available. Since the information consists of independent random draws, such a respondent should not be biased to select travel options which were presented with a specific type of information more often. The opposite of this situation is a respondent who has a very strong preference for a certain type of information and will pick a travel option that is presented with the respondents preferred type of information. We will define the *sensitivity to information*, i.e. si_j, in such a way that the value of 0 corresponds with the first situation and a value of 1 corresponds to the second situation. The formal definition of si_j we are using is as follows:

$$\mathrm{si}_j(C, k) := \begin{cases} \dfrac{\max\limits_{i \in \{0,1,2\}} \mathrm{ic}_j(C, k, i) - \lceil \# PT(C, k)/3 \rceil}{\lfloor 2\# PT(C, k)/3 \rfloor} & \text{if } \# PT(C, k) > 1 \\ 0 & \text{otherwise} \end{cases}$$

In order to detect a tendency of a respondent to repeat a choice for a number of rounds, it is interesting to consider the longest streak of a single choice observed with the choice sequence of a certain individual. If the sequence contains a very long streak this gives an indication that the respondent feels a kind of loyalty toward a certain travel option, while a respondent with many short streaks is more focused toward experimentation. In order to compute the length of the maximum streak, we need to take the maximum over the length of a number of streaks. This can either be expressed using an iterative or recursive procedure. We present the recursive procedure with the variable b as a state variable for the length of a streak.

$$\mathrm{maxstr}'(C, k, b) := \begin{cases} b & \text{if } k = 1 \\ \max\{\mathrm{maxstr}'(C, k-1, 1), b\} & \text{if } c_k \neq c_{k-1} \\ \mathrm{maxstr}'(C, k-1, b+1) & \text{otherwise} \end{cases}$$

We can express the *maximum streak measure* maxstr in the following way.

$$\mathrm{maxstr}(C, k) := \frac{1}{k} \mathrm{maxstr}'(C, k, 1)$$

A different driver for behavior can be the delay witnessed the previous day. Although it seems likely that a delay is included in the satisfaction score reported by the respondent, delays are usually already recorded by public transport operators. It is thus interesting to also consider it a possible driver of switching behavior. The *delay responsive switching* measure dr is defined in a very similar way as the other reactive switching measures, by counting how often a switch occurred if the delay was higher than the average observed delay and no switch occurred if the delay was lower than average. Mathematically, this can be computed by the following expression:

$$\mathrm{dr}_j(C,k) := \frac{1}{k}\left(\sum_{r=1}^{k}[\mathrm{trd}_j(r,c_r) \geqslant \mathrm{avg}_{\mathrm{trd}_j}(r)]x_r + \sum_{r=1}^{k}[\mathrm{trd}_j(r,c_r) < \mathrm{avg}_{\mathrm{trd}_j}(r)](1-x_r)\right)$$

where $\mathrm{avg}_{\mathrm{trd}_j}(C,k)$ is the average delay, computed as follows:

$$\mathrm{avg}_{\mathrm{trd}_j}(C,k) := \frac{1}{k}\sum_{r=1}^{k}\mathrm{trd}_j(r,c_r)$$

Since all our measures are defined as a function that takes some vector and a round k as an argument and produces a number within $[0,1]$, we can easily compare a choice vector of the first and second phase for a certain individual. For this we define the *phase comparison function* which compares a choice vector C for the first phase and a choice vector C' for the second phase, according to one of our behavioral measures f, in the following way:

$$\Phi_f(C,C',k) := f(C',k) - f(C,k)$$

Since $f : \mathcal{C} \times \mathbb{Z} \to [0,1]$, we know that $\Phi_f : \mathcal{C} \times \mathcal{C} \times \mathbb{Z} \to [-1,1]$.

3.5 Results

In this section we discuss the results and findings based on applying our three step methodology to the collected data.

3.5.1 Descriptive Data of Responses

The number of completed quasi-experiments is 515. The distribution of the choices during the first and second round is presented in Figures 3.3a and 3.3b. We can observe that the two fast trains, departing at 7:53 and 7:38 are the most popular choices, the later train being more popular. The car is also a popular choice, but not as popular as the 7:53 train. These choices seem very reasonable, as one is likely to arrive on time for work, while minimizing the amount of time spent traveling. The picture for the second phase is clearly not as smooth as the picture of the first case, giving already a strong indication that the crowding indicators influence the choice behavior of the respondents.

In Figures 3.3c and 3.3d the number of respondents who switch from one travel option in the previous round to another is reported (as such we have no results for round 1, where there is no previous round). During the first phase, the number of switching respondents is typically a little below half. During the second phase this number increases and is also fluctuating more heavily than during the first phase.

3.5.2 Behavioral Measures

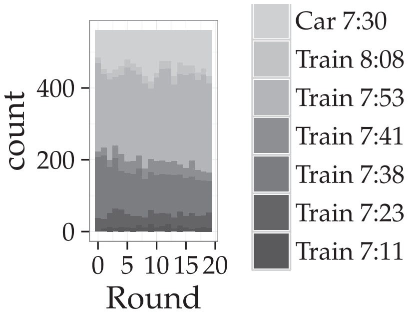
Descriptive statistics of the measures can be found in Table 3.4. The descriptive statistics show some differences between the first and second phase, most notably a major increase in the switching probability, on average by 0.18 and an increased of the satisfaction by 0.04 (on a $[0, 1]$ scale).

We also consider the observed measures for the different groups, and have performed comparisons between these groups. The complete results can be found in Appendix B.2. The measured values for different groups are presented in Tables B.1, B.2 and B.3. The summary of these comparisons can be found in Table 3.5. We discuss the most notable results for the different groups:

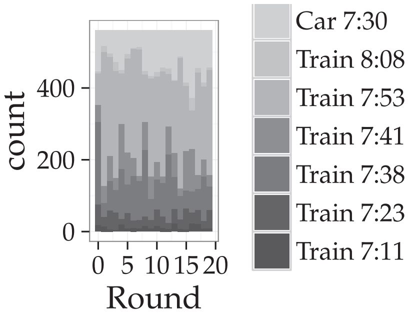
Information: During the first phase of the experiment, the accuracy of information has no impact at all. This is expected, since no information is being communicated during the first phase. During the second phase, the accuracy of information has an impact on the crowding avoidance, the time of the last switch, the switching probability, the tendency to switch when dissatisfied, and the average satisfaction. The availability of accurate information influences all measured aspects of behavior when considering the introduction of information during the second phase. Surprisingly the effect on the sensitivity to information is not significant. It can be expected that a poorer quality of information will make people less sensitive to the information. Although the sign of the observed difference matches this intuition, our evidence is inconclusive.

Disruptions: The occurrence of disruptions has a relatively mild impact on the behavior. Most notably, the lack of disruptions leads to a higher average satisfaction and a lower sensitivity to information.

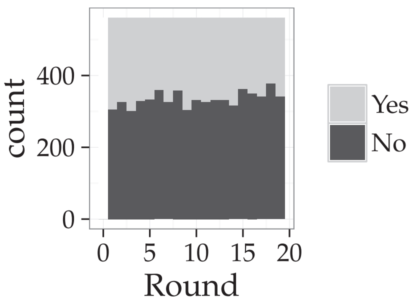
Crowding: Reactive crowding has a notable influence on the behavior. The biggest impact is observed related to the last switch; in case it is possible to obtain lower crowding levels by switching, people tend to keep switching for a longer amount of time. The reactive crowding manipulation also yields lower satisfaction. Crowding Avoidance (ca) is lower in case the crowding is reactive, which can be explained based on the dynamic. If crowding was low during the previous round, it is likely to become high during the next round when the reactive crowding is active. When individuals become aware of this, it seems reasonable



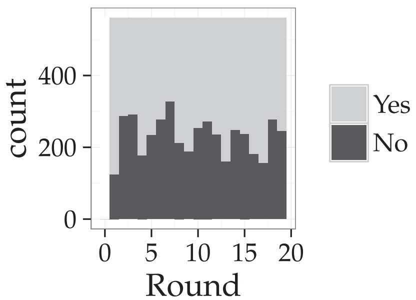
(a) Distribution of choices observed during phase 1



(b) Distribution of choices observed during phase 2



(c) Number of switches during phase 1



(d) Number of switches during phase 2

Figure 3.3: Descriptive statistics of our data set. The top two figures show how many respondents (the x-axis) choose each of the different travel options (the colors in the legend) in each of the rounds (the y-axis). The bottom two figures show how many passengers have switched compared to the previous round, for each of the rounds.

Table 3.4: Table of Summary Statistics over the full population of 515 respondents. The top part shows the different behavioral measures during the first phase of the experiment (i.e. without a crowding indicator). The middle part shows the behavioral measures for the second phase of the experiment (i.e. with a crowding indicator). The bottom part shows the distribution of change of the behavioral measures from the first to second phase for each individual.

			Percentiles				
Statistic	Mean	St. Dev.	Min.	25%	Median	75%	Max.
ca-1	0.34	0.19	0.00	0.20	0.33	0.46	0.92
ls-1	0.65	0.38	0.00	0.29	0.84	0.95	0.95
mtsc_2-1	0.41	0.38	0.00	0.00	0.42	0.79	0.89
sp-1	0.41	0.31	0.00	0.11	0.37	0.68	1.00
rsc-1	0.41	0.31	0.00	0.12	0.40	0.67	1.00
rsca-1	0.59	0.17	0.00	0.50	0.61	0.67	1.00
avg_S-1	0.56	0.08	0.30	0.51	0.56	0.62	0.79
maxstr-1	0.45	0.33	0.05	0.16	0.32	0.76	1.00
dr-1	0.52	0.15	0.11	0.42	0.53	0.63	0.79
ca-2	0.33	0.23	0.00	0.12	0.33	0.50	0.91
ls-2	0.83	0.25	0.00	0.84	0.95	0.95	0.95
mtsc_2-2	0.55	0.36	0.00	0.05	0.74	0.84	0.89
sp-2	0.59	0.27	0.00	0.42	0.63	0.79	1.00
rsc-2	0.59	0.27	0.00	0.41	0.62	0.80	1.00
rsca-2	0.58	0.16	0.17	0.50	0.56	0.67	1.06
avg_S-2	0.60	0.10	0.29	0.54	0.61	0.67	0.83
maxstr-2	0.28	0.24	0.05	0.11	0.21	0.37	1.00
dr-2	0.47	0.13	0.11	0.37	0.47	0.58	0.79
si-2	0.40	0.23	0.00	0.25	0.40	0.55	1.00
Φ_{ca}	−0.01	0.26	−0.67	−0.18	0.00	0.19	0.75
Φ_{ls}	0.18	0.36	−0.95	0.00	0.00	0.32	0.95
Φ_{mtsc_2}	0.14	0.39	−0.89	0.00	0.00	0.32	0.89
Φ_{sp}	0.18	0.30	−0.95	0.00	0.16	0.37	0.95
Φ_{rsc}	0.18	0.30	−1.00	−0.004	0.16	0.36	1.00
Φ_{rsca}	−0.01	0.22	−0.72	−0.11	0.00	0.11	0.78
Φ_{avg_S}	0.04	0.10	−0.23	−0.03	0.04	0.11	0.47
Φ_{maxstr}	−0.17	0.29	−0.89	−0.32	−0.11	0.00	0.95
Φ_{dr}	−0.05	0.17	−0.58	−0.16	−0.05	0.05	0.58

Table 3.5: Table indicating for which behavioral measures a significant impact of the manipulations (represented by the columns) and the introduction of information in the second phase (represented by the Φ values) was found. For each split of the population based on a manipulation, it is indicated whether there was a significant difference ($p \leqslant 0.01$) and how large the observed difference was (+ indicates an increase of less than 0.1, while ++ indicates a larger increase). The comparison is summarized for the first (1) and second (2) phase and the pair-wise differences between the two phases (Φ)

	Information			Disruptions			Crowding		
	1	2	Φ	1	2	Φ	1	2	Φ
ca		++	++		+		-	-	
ls			- -				++	+	
$mtsc_2$			- -						
sp		- -	- -					++	++
rsc		- -	- -					++	++
rsca		++	+	-			-	-	
avg_S		-	-	+	+		-		+
maxstr			++				-	- -	
dr		+	+		-	-		-	
si	n.a.		n.a.	n.a.	-	n.a.	n.a.	-	n.a.

that they switch more often in case the crowding is low. In the situation where crowding is reactive, it is also observed that respondents are less sensitive to information.

3.5.3 Behavioral Clusters

For the third part of our analysis, we will cluster the dataset. Based on a scree-plot, we observed that the best number of clusters is 4 for both the first and second phase data. We run the k-means algorithm on the data for the first and second phase. The cluster means and cluster sizes are reported in Table 3.7.

Based on this table, we observe the following types of clusters which we will interpret as follows:

The stoic (st): The stoic has a very short period of experimentation at the beginning of the experiment and then sticks with a single choice for the remainder of the rounds. The mean of the stoic cluster in the first and second phase of the experiment is very similar. The stoics are fewer in number during the second

Table 3.6: Transitions between clusters comparing phase 1 against phase 2

	Phase 2			
Phase 1	st	du	ms	hs
st	0.25	0.31	0.27	0.17
du	0.05	0.31	0.42	0.22
ms	0.02	0.15	0.43	0.41
hs	0.02	0.07	0.45	0.46

Table 3.7: Results of the three clustering runs. For the first and second phase, four clusters were chosen. The centroid values for each cluster are presented, as well as the number of observations with the clustering and the fraction this comprises of the entire population.

	Clusters Phase 1				Clusters Phase 2			
	st	du	ms	hs	st	du	ms	hs
ca	0.19	0.29	0.41	0.44	0.17	0.36	0.48	0.18
ls	0.04	0.78	0.83	0.93	0.06	0.85	0.89	0.94
mtsc_2	0.01	0.06	0.65	0.82	0.01	0.08	0.77	0.72
sp	0.04	0.28	0.44	0.80	0.04	0.41	0.58	0.86
rsc	0.04	0.29	0.45	0.80	0.05	0.41	0.57	0.87
rsca	0.56	0.61	0.65	0.56	0.65	0.59	0.61	0.52
avg_S	0.58	0.56	0.54	0.56	0.60	0.59	0.57	0.65
maxstr	0.94	0.47	0.30	0.14	0.92	0.36	0.23	0.12
dr	0.64	0.57	0.51	0.38	0.65	0.53	0.48	0.38
Size	130	110	126	149	43	105	202	165
Fraction	0.25	0.21	0.25	0.29	0.08	0.20	0.40	0.32

phase of the experiment, which suggests that the availability of information leads to more variation in choices made. The main reason for stoics to switch seems to be a disruption, as their dr value is a lot higher than the ca or rsc values.

The dualist (du): The dualist has a short period of experimentation at the beginning of the experiment and then keeps switching between two alternative choices for the remainder of the experiment. As such, a dualist has a very low value for mtsc_2, but a high value for the last switch ls. The number of dualists is comparable during the first and the second phase of the experiment. During the second phase of the experiment, the dualists are more likely to switch than during the first phase of the experiment.

The moderate switcher (ms): The final two types, the moderate and the heavy switcher, are similar in regard to the fact that they consider more than two alternative choices for a relatively long time (more than 75% of the rounds). Their main difference is that the moderate switcher switches in 58% of the cases on average. During the second phase of the experiment, the moderate switcher also has a higher crowd avoidance than the heavy switcher.

The heavy switcher (hs): The heavy switcher keeps switching between more than two alternatives for a long time. When compared to the moderate switcher, the reactive switching coefficient is larger in case of the heavy switcher, and thus the satisfaction score seems to be a good predictor of additional switches. It is also remarkable that during the second phase, the crowd avoidance is a lot lower for a heavy switcher than for a moderate switcher, while they have similar levels during the first phase.

It is worthwhile to consider whether these different cluster types appear proportionally when we split the dataset according to our experimental manipulations. We have computed these appearances for both the first and second phase and have applied an exact binomial test to see whether the observed distributions are significantly different from a random distribution. The results are reported in Table 3.8.

The following observations are significant at the $p \leqslant 0.01$ level.

- During the first phase of the experiment, the occurrence of large disruptions leads to fewer stoics.
- During the second phase of the experiment, accurate information leads to a significantly larger number of heavy switchers, while inaccurate information leads to a significantly larger number of moderate switchers.
- During the second phase, reactive crowding leads to more heavy switchers. This suggests that respondents learn that in the setting of reactive crowding it pays off to switch more often.

Table 3.8: Distribution of Treatments over Clusters for both phases. The p-value indicates the probability that the distribution of the two treatments within a behavioral group occurs by chance. Significant results at the $p < 0.01$ level are displayed in a bold font.

	Information			Disruptions			Reactive Cap.		
Cluster	0	1	p-value	0	1	p-value	0	1	p-value
Phase 1									
st	73	57	0.09	**49**	**81**	0.003	78	52	0.014
du	60	50	0.2	58	52	0.31	53	57	0.39
ms	64	62	0.46	75	51	0.02	56	70	0.123
hs	68	81	0.16	73	76	0.43	63	86	0.035
Phase 2									
st	18	25	0.18	15	28	0.033	28	15	0.033
du	45	60	0.08	53	52	0.5	64	41	0.016
ms	**61**	**141**	$9 \cdot 10^{-9}$	103	99	0.42	117	85	0.014
hs	**141**	**24**	$1.1 \cdot 10^{-21}$	84	81	0.45	**41**	**124**	$3.4 \cdot 10^{-11}$

Finally, let us consider to which extent the respondents jump between the clusters from the first to the second phase. For this we can produce a Markov-style transition matrix, where we count the number of respondents who jumped from one cluster to another between the first and second phase and divide it by the total number of respondents belonging to the first cluster in the first phase. This yields the transition matrix presented in Table 3.6

First, it can be observed that the stoics in the second phase of the experiment are likely to also have been stoics during the first phase. Furthermore, when stoics start switching more often during the second phase, they are more likely to become dualists or moderate switchers than heavy switchers. The behavioral measures of these types are simple enough to be able to deduce these from the smart card data of commuters, which makes the findings of practical importance for public transport operators.

Dualists during the first phase are most likely to become moderate switchers or remain dualists when crowding indicators are introduced in the second phase. Moderate Switchers and Heavy Switchers change roles in almost equal proportions in the second phase and they are not likely to become dualists or stoics.

3.6 Discussion and Conclusion

We have conducted a quasi-experiment based on a crowding scenario in public transport. We have obtained time-choice data from 515 respondents, as well as their satisfaction with the reported outcomes. Public transport operators can derive similar data from smart card data and rolling stock allocations. Satisfaction data can either be collected through smart phone apps, or estimated using utility models, which are in widespread use already.

We have been able to find a number of effects caused by our experimental manipulations. First, the introduction of information has an impact on the behavior of the respondents, as they switch more often and are slightly more satisfied on average. When random information leads to lower satisfaction than accurate information, it also leads to more crowd avoidance and a lower switching probability. Secondly, the absence of disruptions leads to higher satisfaction. Finally, increasing the crowding of the previous choice leads to longer periods of switching.

Based on our measures, we are able to distinguish four types of behavior: Stoics, who decide early and stay with their decision; dualists, who decide on two travel options early and keep switching between these two; and the moderate and heavy switchers. We observe that the occurrence of large disruptions leads to fewer stoics during the first phase of the experiment. During the second phase of the experiment the availability of accurate information leads to more heavy switchers, while random information leads to more moderate switchers. The reactive crowding incentivizes switching behavior with regard to crowding and we indeed observe that purely random crowding leads to fewer heavy switchers than the case of reactive crowding, but this difference is only significant during the second phase of the experiment, when information is made available.

This implies that crowding influences behavior and more so when crowding information is provided to the respondents. The availability of information also has an impact on behavior, as well as the quality of the information provided. The fact that the crowding regime (random or reactive) leads to a change in behavior when information is available, suggests that respondents are able to discover that crowding is reactive and that at least some of them adapt their behavior accordingly.

These insights are useful for the development of an agent-based simulation model, as it offers four behavioral types which can be mimicked by such agents. For the stoic it makes the most sense to pick a preferred travel option once and stick to it. For the other behavioral profiles, we propose two different ways to implement them in a simulation study:

Randomized Behavior: There is always a certain probability that an agent switches from one choice to another. These probabilities can be influenced by events that occurred during the previous round. The probability based behavioral measures can be used to determine the probability that an individual will switch given a set of circumstances. A randomized draw can be used to decide whether or not an agent should switch based on this probability. In case of a switch, the agent must also decide which new option will be chosen. In case of the dualist this is an easy choice. In case of the moderate and heavy switchers population wide probabilities can be used to decide the new choice.

Deterministic Behavior: The occurrence of an event (e.g. too crowded, a disruption) will always or never trigger a switch; a deterministic agent does not utilize random draws to make decisions. The events in which a switch will occur will be selected randomly a priori using the behavioral parameters.

Furthermore, these behavioral profiles can also help in forecasting future passenger choices. For stoics the strategy to predict that the previous choice will be repeated will be very accurate. For dualists a forecasting strategy that focuses on when a switch will occur will be an effective forecasting strategy. Since both groups already make up quite a large part of the population, treating these groups separately will likely result in a more accurate forecasting model.

Future research aims to combine these findings with our existing simulation framework from Chapter chp:eftg in order to develop an agent-based simulation model. Operators should be able to apply the same process to smart card data in order to create a useful tool for crowding, in which rolling stock planning and information policies can be evaluated for different behavioral groups. As a result such an agent-based model can be a valuable tool for evaluating different strategies for capacity decisions in a dynamic setting where passengers react to crowding and information.

Part II

Analysis of Smart Card Data for Public Transport Design

4 Recognizing Demand Patterns from Smart Card Data for Agent-Based Micro-simulation of Public Transport

Co-authors : Milan Lovric, Ting Li, Evelien van der Hurk, Leo Kroon and Peter Vervest
This chapter is an improved version of a paper has been accepted for inclusion in the proceedings after peer review by the program committee of the workshop on *Agents in Traffic and Transportation* of the AAMAS 2015.

4.1 Introduction

In public transport systems without seat reservations, the question of how fluctuating demand can be serviced in a cost-efficient way poses a major challenge. Peaks in demand have a high toll on the costs, since they dictate the required amount of staff and the number of vehicles, while vehicles that are almost empty generate a net loss for the operator. Tools that allow the public transport operator to evaluate the effects of operational and strategic decisions on costs and demand are therefore vital to achieve the goal of improving the service quality and financial performance. Most of the tools used in practice model passengers as aggregate flows rather than individuals, either because detailed data is not available or to reduce the complexity the decision maker has to face. During recent years, smart card systems have been introduced that log all movements of individual passengers through the systems. This gives a lot of detailed data that was previously unavailable. However, given the body of research related to smart card data, it has been observed that incorporating such data into the tools and models used for decision making is a non-trivial task (Pelletier et al., 2011).

A promising approach is agent-based micro-simulation. In such a simulation, individual passengers and vehicles are modeled through agents that interact with the public transport system according to their individual goals, consisting of various activities that need to be performed at various locations. In this chapter, we use the MATSim simulation package (Horni et al., 2016) which has an active user-base and has been applied to a number of large scale scenarios. Within MATSim, all agents try to adapt their plans in such a way that their utility is improved. The simulation runs until there is no significant improvement within the agent population, i.e. until the population reaches an approximate equilibrium.

The major issue in generating an agent population from real life observations is the question how we can prevent agents to divert from this equilibrium in an unrealistic way, without restricting the agents in such a way that their only preference is to replicate the observed state. Furthermore smart card data is merely a proxy of the actual activities that individuals perform and therefore additional assumptions are required to build a model of demand driven by activities instead of fixed flows.

We will limit our field of application to the study of *revenue management*. In revenue management (Talluri and Van Ryzin, 2005) the goal is to control demand by adapting our pricing strategy in such a way that we get a better match between the available capacity and the demand emerging from the population. Our population can try to adapt to our pricing strategy by shifting the time at which they travel. We will study how the population reacts to an off-peak discount, but we believe that our approach is suitable for many similar case studies.

When generating our agent population, we run into the problem that the number of observed journeys differs a lot between individual passengers. We solve this problem by combining three types of demand that we can detect in our smart card dataset: *trip-based*, *tour-based* and *pattern-based* demand. Our first goal is to show how we can efficiently generate the agent population from our smart card data using these three demand models. Our next goal is to discuss how we can experiment with different parameters for the demand models to study revenue management. The final goal is to discuss our results and how we can improve our methods in the future.

The remainder of this chapter is organized as follows: In Section 4.2, we discuss prior literature and related work. In Section 4.3, we discuss smart card datasets in general and our dataset in particular. Section 4.4 addresses the modeling of demand, based on the smart card dataset. In Section 4.5, we discuss the simulation and our experimental setup. We present the results of our experiments in Section 4.6. Finally, we discuss our results and opportunities for extensions of our approach in Section 4.7.

4.2 Related Work

In recent years, smart card ticketing systems have attracted notable attention from the research community. A literature review on the use of smart card data in public transport is given by Pelletier et al. (2011). They divide the research into three categories: Strategic-level studies, tactical-level studies and operational level studies. Since some of the public transport systems only work with check-ins, part of the literature focuses on estimating the destination of passengers given their check-in location and time, for example in Trépanier et al. (2007). Some literature describes how the behavior of passengers can be analyzed. A notable example is Morency et al. (2007), where spatial and temporal variations are measured across different types of cards. However, the literature review by Pelletier et al. (2011) contains not a single reference to the use of smart card data within a simulation context. Moreover, their conclusion contains the following quote:

> For the mass of data available on individual trips, new modeling methods will be needed, such as the Totally Disaggregate Approach, because classical models cannot be used at a such detailed level of resolution. [...] It will then be possible to calibrate individual base models from these large datasets.
> (Pelletier et al., 2011) (page 566)

In the simulation of road traffic, microscopic simulation models have been a topic for quite some years. In the 1990's, it was mostly a topic studied as a field of application for super computers (Cameron and Duncan, 1996). With the increase of computing power, more applications emerged in the 2000's, including Treiber et al. (2000). With the introduction of MATSim (Horni et al., 2016), we saw a rise in literature related to micro-simulation. MATSim has been applied to some very large scale scenarios, including simulations of Berlin (Rommel, 2007) and Zürich (Meister et al., 2010), both including more than a million individual travelers. At some point MATSim was expanded from the simulation of road traffic, to the simulation of public transport (Rieser, 2010). The website of the project contains a list with the most important publications related to the project and is updated regularly.

The kind of microscopic demand which is fundamental in the design of MATSim, is called *activity-based* demand (Bowman and Ben-Akiva, 2001) and was already discussed in the context of micro-simulation by Miller (1997). This is an approach where travel demand is modeled by means of the activities the individual travelers want to perform over the day. One way to record activities of individual travelers is by using surveys, for example Axhausen et al. (2002). In recent studies, census data was used to perform synthesis of activity based demand (Axhausen et al., 2008). A survey on this approach to demand generation is given by Müller and Axhausen (2010).

Apart from modeling the activity patterns of travelers, much research regarding the behavior of travelers has been performed, resulting in many sophisticated methods. Most notably, we must mention the field of discrete choice modeling (Ben-Akiva and Lerman, 1985), since it has spawned a lot of research within the domain of transportation. One of the main tools within discrete choice modeling is the stated-choice survey, where respondents have to select their preferred alternative from different available alternatives.

A comprehensive textbook on the research on and practice related to revenue management is Talluri and Van Ryzin (2005). The focus of research related to revenue management has been on systems where reservations are made in advance. In our setting, however, we do not have a mechanism where we can decide whether we accept new customers. This is different from, for example, long distance trains and the airline industry where tickets are always bought in advance. An example of a study related to revenue management in a comparable railway setting is Link (2004). This study shows some of the difficulties in applying revenue management within our context. An example of a successful application for long distance trains with seat reservations is Ben-Khedher et al. (1998).

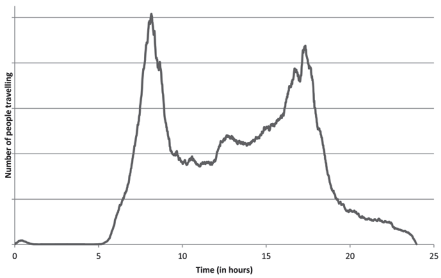
4.3 Smart card data

During recent years, the Dutch smart card, called "*OV-chipkaart*" was introduced as a cross-operator travel product. Starting from 2009, the smart card was made the mandatory product of travel in major Dutch cities, such as Amsterdam and Rotterdam, replacing paper tickets. One of the unique features of the Dutch system is that passengers have to check in and check out with the smart card in all modes of travel, including railways.

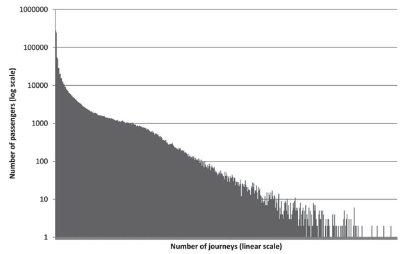
We use data collected from smart card usage over the course of four months from a major public transport operator in The Netherlands. During this period, the only available tickets were different smart card products. The transactions in our dataset denote either a check-in or check-out in a vehicle or on a platform. Moreover the smart card data contains the mode of travel, the unique id of the chip on the smart card (which we call the media id), the time stamp of the transaction (in seconds) and the location of the transaction. Due to the sensitivity of the data for the operator and privacy concerns for the passengers, we only show relative numbers and figures.

We prepared our raw dataset of almost 60 million transactions in such a way that we could process each transaction sequentially. We had to split up the dataset into separate chunks, using a round robin approach to assign media id's we had not seen before to a fixed chunk for that id. Afterward, we sorted the separate chunks on media id and time stamp in main memory. We combined the results into a single dataset. While processing this set sequentially, we would be sure to encounter all transactions belonging to a certain media id together, with increasing time stamps.

After sorting the dataset, we can link check-ins and check-outs. People forgetting to check in or check out gives some inconsistencies in the dataset. It is relatively easy to filter these inconsistencies out, by assuming that a consecutive check-in and check-out belong together. This is reasonable, since the system has a maximum amount of time after which a check-in becomes invalid. After this linking step we know all the trips made by the passenger. Since the passengers have to check in and check out in each vehicle, we have separate trips when the passenger makes a transfer on his journey. Another preprocessing step is to link consecutive trips that are close in time to each other into journeys. This yields our main dataset. Figure 4.1a shows the numbers of unique passengers traveling over the course of a typical weekday. Figure 4.1b shows a histogram describing how many journeys were made with a single smart card. As we can see, most of the smart cards have made only a relatively low number of journeys, but there are plenty of passengers with many journeys.



(a) Demand histogram of a weekday



(b) Histogram of the number of passengers that made a certain number of journeys within 4 months

Figure 4.1: Demand as observed in the smart card dataset

4.4 Demand modeling

When it comes to demand modeling for the simulation of public transport, a traditional approach is to use origin-destination matrices estimated from sources such as census data and manual counts of the number of passengers in some sampled vehicles (Müller and Axhausen, 2010). The main drawback of this approach is that it becomes very difficult and expensive to measure the exact progression of passenger flows over the day. With smart-card data, we know the origin, destination and exact time of travel of each individual travel, which allows for new opportunities with respect to measuring these flows.

Regarding flows of passengers in the network, we can take different approaches. The basic approach is to consider a flow through the network as a set of journeys: Passengers who travel from a certain origin to a certain destination at a certain time. We will refer to this approach to demand as *trip-based* demand. However, in many case there will be passengers who travel multiple times within the same day. In many of these cases, their consecutive journeys combine to a tour from origin to origin, with some intermediate destinations. In such cases, events happening at one of the intermediate destinations, will also influence the events in the remainder of the tour. Since our goal is to model individual passengers instead of aggregated flows, these tours contain valuable information. We will refer to this approach to demand as *tour-based* demand.

In activity-based micro-simulation, each individual traveler can be represented by an agent and this approach thus allows for microscopic analysis of a public transport system. The drawback is that we need a lot of information to model these agents. Even if we assume that all activities take place at a station, not all required information is available in the smart card data. The smart card data tells us *where*, *when* and *how* people travel, but it does not tell us *why* people travel, which is something that is vital to activity-based demand modeling.

Traditional methods use various statistical methods and interpolation techniques to fill the gaps of unknown information, in order to be able to simulate a public transport system. We can apply such an approach to the smart card data as well: We use the information which is available, such as location, modality and time of travel as much as possible and fill the gaps of information using estimation methods.

We introduce an approach that goes beyond the notion of tour-based demand, but does not yet reach the precision of activity-based demand, as *pattern-based* demand. In a broad sense, we define pattern-based demand as demand produced by activities of such a nature that certain patterns will emerge in the travel behavior of passengers who perform the activity routinely. A typical example of such an activity is working, since people usually work at regular times at a certain location. Other types of activities are education (which is usually bound to a schedule that may or may not change regularly), a periodic visit to family members and visiting sports events. In this chapter, we will focus on patterns generated from working activities, since we believe that these will be most easy to recognize. In addition to this, we consider educational activities with a fixed schedule as a working activity, since the implications for the temporal flexibility of a passenger are usually similar. To summarize, we have:

Trip-based demand Demand with only a single journey

Tour-based demand Demand consisting of a tour of journeys, with consecutive arrivals and departures at the same station. Also, the first and last station are equal.

Pattern-based demand Demand that exhibits a recurring pattern, produced by a regular underlying behavior of the passenger.

4.4.1 Detecting customer patterns

Commuters usually live and work at the same place. This leaves patterns of frequent *home* $\rightarrow$ *work* $\rightarrow$ *home* journeys in the smart card data. We can scan consecutive journeys for these patterns. This way we can derive an activity profile for a customer. For the sake of convenience, we limit ourselves to the class of activity profiles described in the following definition:

Definition 1. *Activity Profile*
An activity profile is a tuple $(l, b_{\text{pref}}, e_{\text{pref}}, \delta_b, \delta_e)$ *where*

- *The activity takes place at location* l
- *The preferred starting time of the activity is* b_{pref}
- *The activity should not start before* $b_{pref} - \delta_b$ *and not after* $b_{pref} + \delta_b$
- *The preferred ending time of the activity is* e_{pref}
- *The activity will not end before* $e_{pref} - \delta_e$ *and not after* $e_{pref} + \delta_e$
- *The preferred duration of the activity is* $e_{pref} - b_{pref}$

Now for each passenger, we try to decide whether he is commuting and what his home and working stations are. To do this, we have to make a few assumptions.

1. We assume that somebody who is commuting travels a lot. Therefore, if the number of times traveled in the considered dataset is not above a certain threshold (which should be chosen according to the length of the time period under consideration), we conclude that the passenger is not a commuter.
2. We assume that a commuter has a fixed home and a fixed location of work and that the stations associated with these locations will be the two most frequently visited stations. To be sure these frequent stations are visited more frequently than other stations, we define thresholds for the number of times they should occur.
3. We assume that, if we include weekends, someone will spend more time at home than at work. Since we can measure the time between a consecutive arrival and a departure from a station, we classify the station where most time is spent as the home station.
4. We assume flexibility in time of travel and the length of the working activity is represented by a certain amount of variation in their travel times between their home and working stations.

We use the first assumption to decide whether we try to recognize a pattern for a certain passenger at all. The second and third assumptions can be used to recognize a passengers home station and working station. Finally, we use the fourth assumption to model the flexibility of a passenger based on this variance. These assumptions results in the procedure shown in Algorithm 2.

It is not difficult to see that each of the steps can be performed in time linear with respect to the set of journeys J, except for line 4, where we have to calculate frequency statistics. To take the first and second most frequent station, we can sort the stations based on their frequencies. Since at most $O(n)$ stations occur in J, this gives a $O(n \log n)$ time bound. In Skiena (2008), it is discussed that this selection problem takes $O(n \log n)$ time in general. Since there are no loops in the algorithm, we may conclude that it runs in $O(n \log n)$ time for a single passenger with n journeys.

Algorithm 2: Algorithm for the detection of customer patterns based on the assumption that most time is spent at the home location and home and work locations occur as most frequent in the smart card data.

Parameters: A minimum sample size θ, thresholds t_0 and t_1 with $0 < t_0, t_1 \leqslant 1$
Input : A set J of n journeys of a single passenger
Output : A home station s and a pattern $(t, b_{pref}, e_{pref}, \delta_b, \delta_e)$ that describes a working activity profile as defined in Definition 1

if $n < \theta$ **then**
    Conclude there is no valid pattern
end
Find two stations a, b with maximal frequency as start or endpoint over the journeys in J ;
Denote n_a, n_b as number of journeys that have a or b as a start or endpoint, n as the total number of journeys in J ;
if $\neg(n_a \geqslant t_0 n \wedge n_b \geqslant t_1 n)$ **then**
    conclude there is no valid pattern
else
    $\Delta_a :=$ average time difference between consecutive (a, b) and (b, a) journeys ;
    $\Delta_b :=$ average time difference between consecutive (b, a) and (a, b) journeys ;
    if $\Delta_a \geqslant \Delta_b$ **then**
        $s := a; t := b$;
    else
        $s := b; t := a$;
    end
end
Take the average arrival time of (s, t) journeys as preferred starting time b_{pref} ;
Take the average departure time of (t, s) journeys as preferred ending time e_{pref} ;
Take the standard deviation of (s, t) arrival times as the start time flexibility δ_b ;
Take the standard deviation of (t, s) departure times as the end time flexibility δ_e ;
return $s, (t, b_{pref}, e_{pref}, \delta_b, \delta_e)$

4.4.2 Deriving the Agent population

We will now discuss how to derive an agent population from our dataset. In the beginning of Section 4.4, we discussed the difference between *trip-based*, *tour-based* and *pattern-based* demand. Since there are smart cards that are used only once and passengers who have highly irregular travel patterns (because they do not use public transport to commute), we not be able to derive a pattern for each customer and we may not even be able to find a tour in the data for each customer. Therefore, we take a step-wise approach, where we first try to calculate a pattern for a passenger. If this succeeds, we generate demand for this passenger based on the pattern we found. If we fail to find a pattern, we search for a tour and generate tour-based demand by introducing dummy activities at the intermediate stations of the tour. If we even fail to find a tour, we generate trip-based demand by generating agents for each trip the customer made.

We will choose a single day (preferably not during the weekend) to model. We first filter our dataset such that we only retain customers that have traveled on that day. After filtering, we decompose our dataset into three parts: One group contains customers of which we know a lot, one group contains customers of which we have a tour and lastly, one group of customers with a single or unpredictable travel pattern. For each customer, we have to generate an activity plan for the day. We will take a different approach to the generation of plans for each group of customers.

A plan for the day is a list of activity profiles with planned ending times for all activities. There is one exception: The last activity of the agent should be a home activity, which has no ending time. The ending time in the plan of an agent may differ from the ending times in the activity profile: An agent may try to deviate from his preferred time if this gives him an improvement in utility. The planned ending time is exactly what allows the agent to do this. When we start generating plans for our agent population, we initially stick with the preferred ending times from the activity profiles as the planned ending times. For the group of customers for which we have derived a pattern, we can generate a *home* $\rightarrow$ *work* $\rightarrow$ *home* activity plan. For the group of customers for which we only have a tour, we only have a set of locations. For the activity profiles, we can derive a starting and ending time, using the check-out and check-in time at each intermediate station. While it is reasonable to assume the observed arrival and departure time are close to the preferred start and end time of the activity, the smart card data does not tell us how flexible the preferences of the passenger with regard to these times were. For the time being, we decide to select a global value for the δ_b and δ_e of tour-based agents. We take a similar approach with the trip-based customers, where we generate a single agent for each trip. For each journey we observe from u to v, we generate an agent with a *home*$\rightarrow$ *dummy*$\rightarrow$ *home* pattern, where the first home activity should be performed at location u and the dummy and last home activity should be performed at location v. This prodecure for demand generation is further explain in Algorithm 3.

The running time of this algorithm is proportional to the size of the J_c sets. Let us define $n = \sum_{c \in C} |J_c|$. If we define $k = |C|$ as the number of customers and m as the maximum number of journeys for a single customer, we can easily see that $n \leqslant mk$. Lines 1–4 are regular filtering steps, that can be performed by examining each set J_c or by applying the earlier algorithm and can therefore all run in $O(km \log m) = O(n \log m)$ time. The loops in lines 6–19 each iterate at most over k customers and generating the plan for each customer can be done in $O(m)$ time. Therefore, lines 6–19 run in $O(mk) = O(n)$ time as well. Therefore, the whole algorithm runs in $O(n \log m)$ time.

Algorithm 3: Algorithm for the generation of trip, tour and pattern based demand for a single day of the simulation. The journeys of a group of customers are used to generate the demand.

Input : A day d and a set of customers C with for each $c \in C$ their respective set of journeys J_c

Output: An agent population for day d

1 $P := \{p : p \in C, J_c$ contains a journey during day $d\}$;
2 $P_{pat} := \{p : p \in P, J_p$ has a pattern $\}$;
3 $P_{tour} := \{p : p \in P \setminus P_{pat}, J_p$ makes a tour at day $d\}$;
4 $P_{trip} := P \setminus (P_{pat} \cup P_{tour})$;
5 Initialize agent set $A := \emptyset$;
6 **foreach** $p \in P_{pat}$ **do**
7   Generate an agent with a "*home* → *work* → *home*" plan ;
8   Add the agent to A ;
9 **end**
10 **foreach** $p \in P_{tour}$ **do**
11   Generate an agent with a plan containing the tour locations and ending times of p's tour at day d ;
12   Add the agent to A ;
13 **end**
14 **foreach** $p \in P_{trip}$ **do**
15   **foreach** (u, v) *journey traveled by* p *on* d **do**
16   Generate an agent with a "*home* (at u) → *dummy* (at v) → *home* (at v)" plan such that the *dummy* activity starts at the check-out time of the journey ;
17   Add the agent to A ;
18   **end**
19 **end**
20 **return** A

4.5 Simulation

4.5.1 MATSim

For our agent-based simulation, we used the MATSim 0.3.0 software package. To run a MATSim based simulation, we need three ingredients: The agent population, a network describing how vehicles can travel between nodes and a public transport schedule. When we start the simulation, all agents calculate an initial plan. The main loop consists of a replanning and a simulation phase. During the planning phase, each agent can adapt his activity plan. They do so by using certain modules available in MATSim, called mutators. During the simulation phase, all plans are executed and all events related to movements and activities of agents and vehicles are generated. The mutators used by the agents to adapt their plans, can be given individual probabilities. An example of such mutators are the rerouting mutator, that recalculates the fastest route between activities based on the network congestion of the previous day. Another example is the time mutator, that shifts the planned starting and ending times of the activities randomly, while retaining their sequential order.

At some point the mobility simulation of MATSim was extended with support for public transport (Rieser, 2010). This mobility simulation is an extension of the road-traffic simulation. We may define networks of different modalities and the public transport vehicles are modeled as cars with a driver and a lot of space for additional passengers. To generate the required network, we used a list of stations with their geographical locations and the available schedule information for all three modalities. We add the stations as nodes in the network. If there was a vehicle that visited two stations consecutively in the schedule, we added a link between the two stations, with the distance of the link based on Euclidean distance between the two stops. We enforce the vehicles to wait at each stop until their scheduled time of departure. The mobility simulation itself is a discrete event simulation through a queuing network generated from the input network.

MATSim allows us to transfer money from or to an agent, but this mechanism is not triggered automatically. We added a module that imposes fares on the agents. It keeps track of the moments agents enter and exit the vehicles and the distances traveled by the vehicles. The fare of a journey consists of a *base tariff* that is the same for all journeys and a *distance tariff* with a certain fixed amount per meter traveled. An additional aspect is the transfer time: If the check-out time and check-in time of two consecutive journeys is small enough, the agent does not have to pay the *base tariff* a second time. At the end of the simulation of a single day, the accumulated fares are billed to the agent and taken into account during the evaluation of the utility of the executed plan.

θ	pattern	tour	trip
80	26%	32%	42%
120	14%	40%	46%
160	4%	48%	48%
200	1%	50%	49%
∞	0%	50%	50%

Table 4.1: Population distributions for different θ values

4.5.2 Experimental Setup

We processed the smart card data on a desktop PC with a quad-core Intel Core 2 Quad Q6600 processor and 8 GB of RAM running Windows 7 Professional SP1, 64-bit. Since we want our passengers to have their working station in at least half of their journeys and we want their home station to be at least as frequent as their working station, we chose $t_0 = 0.6$ and $t_1 = 0.5$. Prior to our experiments, we generated populations for different values of θ. Since θ determines the minimum number of journeys before we use a smart card to generate pattern based demand, the θ will have an important impact on the generated population as a higher value will push more and more smart card into the tour based or trip based demand. We examined a few possible values of θ, of which the resulting percentages of pattern based, tour based and trip based demand that are being generated are presented in Table 4.1. In the extreme case, we observe that 50% of all smart cards can be used to generate tour-based demand. If we require 80 journeys in the full time periode, roughly 25% of the smart cards can be used to generate pattern based demand. Since this percentage rapidly declines when we increase θ, we use $\theta = 80$, $\theta = 120$ and $\theta = \infty$ in our experiments where we assess the impact of pattern based demand in the simulation. For the tour-based and trip-based demand, we wanted our agents to keep as close as possible to their observed travel time, so we fixed δ_b and δ_b for their activity patterns to 5 minutes. This gives us a total of three different agent populations.

The simulations themselves were performed on a desktop PC with an Intel i7-4770 processor and 16GB of RAM running Windows 10. For our pricing strategy, we took figures inspired by the real world pricing policies. We set the base tariff to 0.75 and the distance based tariff to 0.115. The allowed transfer time is set to 30 minutes. For our experimentation with revenue management policies, we ran each of our populations through the network two times: Once with a single tariff over the full day and once for each discount percentage of 1%, 5% and 40% outside the peak hours (the peak hours are between 7:00–9:00 and 16:00–19:00). This way we can compare the effect of a small discount with the effect of a larger discount. The check-in time determines whether the discount is given, which is similar to various peak-hour pricing schemes used in practice. To allow agents to perform time shifts, we enabled MATSim's time mutator module. Since we run each population against all four pricing policies, we get twelve scenarios for which we run a simulation experiment.

After some preliminary experiments, we saw that the increase of agent-utilities slowed down significantly between the 60th and 100th iteration. To be sure our simulation reached a state that is close to an equilibrium, we ran each simulation up until the 180th iteration.

4.5.3 Scoring, Utilities and Price Elasticities

The standard way to compute the utilities of agents within MATSim is called the Charypar-Nagel scoring function Charypar and Nagel (2005). In this section we consider the basic idea of this scoring function. First and foremost, the scoring function results in monotone non-decreasing utility for the activity when it is performed longer (and strictly increasing in case the activity is available), and monotone non-increasing utility for longer travel times. Thus, agents maximize their utility when they spend their time doing as little traveling as possible while balancing the activity times. Activities are defined to have "*typical durations*" and the scoring function is defined in such a way that the highest utility is obtained when the time spend doing different activities is proportional to their typical durations.

For a typical commuter, two activities are important: The home activity and the work activity. In a regular case, the work activity is available in a certain time window. If the agent arrives to early it has to wait before the activity starts. This scenario results in opportunity loss, as the time spend waiting could also have been used to extend the home utility. As the work activity becomes available there is a window in which it can be started. Starting too late may result in the situation where a good duration of the work activity cannot be obtained, resulting in opportunity loss. The scoring function also adds an additional penalty if an activity starts later than a given time, resulting in even greater decreasing utilities in such situations.

Since we are interested in the impact of time shifting on the utility of the agents, we consider the impact of the departure time x_d in the morning on the utility u_a gained by performing activities. The general structure of this relationship is expressed in Figure 4.2 Although the actual scoring functions work with logarithmic curves, we use a linear representation for the sake of simplifying the discussion. Both logarithmic and linear function are monotonic and continuous in the considered areas, which makes the similar from the perspective of a heuristic search for an equilibrium.

The function presented in Figure 4.2 has the following segments: The first part represents the opportunity loss at the home activity in case the agent arrives too early for the work activity. The second segment represents the fact that the time the activity starts does not matter too much, as long as its duration can be balanced with the other activities. The third segment represents the part where the agent is not late at work, but is unable to perform work long enough to properly balanced it with the home activity. The final segment represents the fact that the agent is late at work and an additional penalty is incurred.

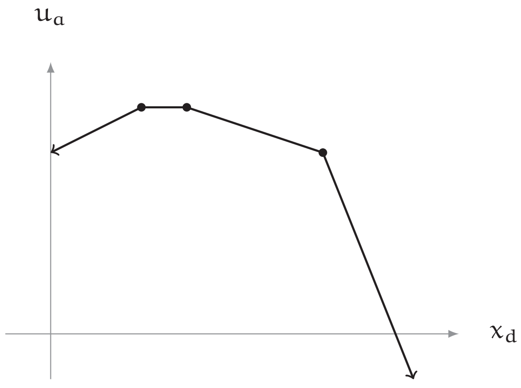


Figure 4.2: Structure of the impact of the departure time in the morning x_d on the utility term of utility gained by performing activities u_a.

The structure of Figure 4.2 is ideal from an optimization perspective: By evaluating small deviations from the departure time, better utilities can be found until the global maximum is reached. As the function is continuous and concave, it is very easy to maximize. This structure changes as soon as we introduce a period discount on the travel fees. In Figure 4.3 we see the structure of the utility based on ticket price u_p given the departure time of the agent in the morning x_d, assuming the fee of the journey is otherwise fixed.

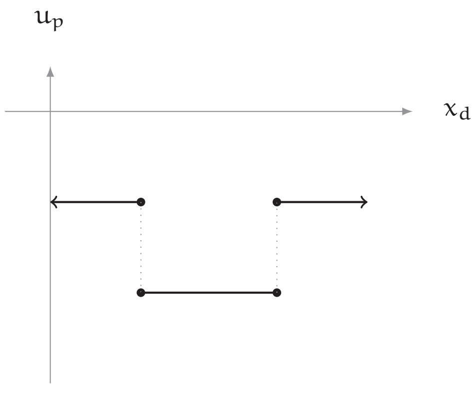


Figure 4.3: Structure of the impact of the departure time x_d on the utility term caused by different ticket price u_p due to the costs of the ticket in case of a markup or discount period, assuming the durations of the activities do not change.

We can clearly see that Figure 4.3 contains a sharp discontinuity. When we add this to the regular scoring function from Figure 4.2 the resulting function is discontinuous and non-concave. The resulting function is a lot harder to optimize by simple heuristics, due to the introduction of local optima.

Another important difference between car traffic and public transport is that public transport is schedule based. This has implications for the utility function, as waiting at a station or bus does not yield any utility. As a result, it is best for agents to arrive at a station or bus stop just before their intended vehicle depart, as waiting at a station results in less time that can be spend performing utilities. The resulting impact of the time of departure x_d on the utility due to waiting for a public transport vehicle to depart u_w is presented in Figure 4.4.

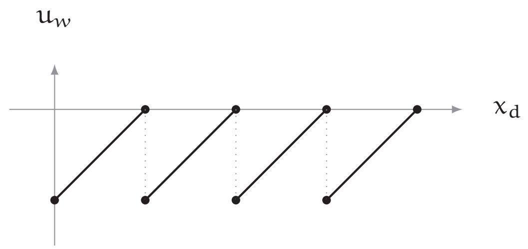


Figure 4.4: Structure of the impact of the departure time x_d on change in utility caused by waiting for the next public transport service u_w.

We can clearly see that the utilities caused by waiting for a public transport vehicle introduce another layer of discontinuity in the utility function. As a result, we can conclude that optimizing utilities by time shifts in case of public transport with off-peak discounts is more complex than for car traffic, as agents have to deal with more discontinuities.

One way to overcome this issue is to use a different time shifting mutator. The default mutator available in MATSim randomly adapts the time of departure. For car traffic, it makes sense that individuals in the real world do something like this: If they arrive for work, they depart a few minutes earlier the next day and check whether that helps. In case of public transport, a more advanced mutator that explicitly takes the time-table and discount periods into account would very likely be beneficial. This information can be used to determine potential local minima and jump to those points, instead of randomly mutating the departure time. Furthermore, many passengers of public transport make use of timetables to plan their journeys, so it can be argued that within the context of public transport such a strategy resembles actual behavior better than the current mutator.

For our simulations we did consider to implement an extension of this scoring function that assigns a personal price sensitivity to each agent, but we decided to keep this is currently a common parameter for all agents. We used 6 and -6 as the (global) coefficients for the performing and traveling utilities and -18 as the coefficient for late arrival. Note that price elasticity is not defined directly in MatSIM, but rather is an emergent property of the utility coefficients defined and the properties of the activities, notably the typical duration of each activity. In some preliminary experiments we found that the simulations are quite sensitive to the typical duration assigned to the home activity. We decided to set this typical duration to twelve hours for all agents. However, it must be noted that it is important to compare the price elasticity that emerges from the model with price elasticities measured in other ways before it can be applied in practice. Two relevant papers that investigate possible ways to calibrate these types of emergent macroscopic properties are Flötteröd et al. (2011) and Flötteröd (2017) of which the latter considers an application to finding a good road pricing policy.

4.6 Results

Generating our one of the agent populations we use for simulation can be done very efficiently from the dataset we discussion in Section 4.3. The average time required to process this full set of 27 million journeys and write the agent population to MATSim input files was on average 107 seconds. As Input/Output seemed to be a signficant bottleneck, it is likely that these times can be further reduced with SSD storage and modern hardware.

Our simulation executed a complete iteration of the mobility simulation in about half a minute. Some additional time was needed for routing computations performed by MATSim and also disk IO due to the dumping of all the plans of the agents after each 10th iteration. A complete run of a single scenario took roughly 95 minutes. The vehicle loadings observed after the first iteration were in all of our six scenarios relatively similar to Figure 4.1a. But at the 180th iteration, we saw notable differences.

During the experiments we measured the number of passengers currently making use of the public transport system (the passenger count) and the revenue generated by the travelling of passengers. We normalized these values by defining the highest observed value as 100. The maximum passenger count and the collected revenue observed during the simulation of each scenario is reported in Table 4.2. The different observed passenger counts for a plain tariff and the 40% discount scenarios are shown in Figure 4.5.

Let us consider the impact of the three levels of θ, i.e. $\theta = 80$, $\theta = 120$ and $\theta = \infty$. Based on Figure 4.5 we can see that for a plain tariff, the morning peak becomes less steep when we increase θ and that it also influences the height of the afternoon peak. In Table 4.2 we can observe that the highest observed passenger count decreases when we move from $\theta = 80$ to $\theta = 120$, but increases when we move from $\theta = 120$ to $\theta = \infty$. However, the revenue increases for both steps. This suggests that the total use of the public transport increases as we include less pattern based demand, but that this also causes subtle shifts in time that results in a non-straightforward behavior of the highest observed peak passenger count.

The scenarios with the high discount of 40% show a different picture. In the case of $\theta = 80$, we can see that the peak demands are decreased by this discount. Especially the morning peak is split up into two lower peaks, one of which lies outside the discount window. As we increase θ to 120 or even ∞, we can see that this split in the morning peak hour becomes less pronounced, but remains visible. Table 4.2 shows that the maximum passenger count increases as we increase θ within the 40% discount scenario, while the same holds for the revenue. This relation is similar for the 5% discount, but at a 1% discount the maximum passenger count for $\theta = 120$ is lower than for $\theta = 80$ and more similar to the scenarios with no discount. This suggests that if the goal is to reduce peak demand, one must carefully evaluate the reaction of the population to a discount. This is also suggested by the percentage point per discount percent, which is close to 0.2 for the $\theta = 80$ case, but varies quite a bit for $\theta = 120$ and $\theta = \infty$. This shows that relation between peak demand and the discount is non-linear in nature.

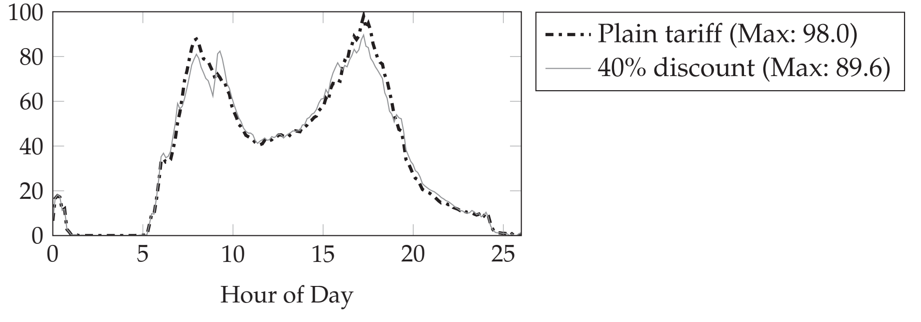
As the $\theta = 80$ has the highest number of agents with pattern based demand, we would expect these agents to have more opportunity to perform a time shift than in the $\theta = \infty$ scenarios. This is also supported by the more pronounced change in peak demand observed during the morning peak for the $\theta = 80$ scenarios. However, the results also suggest that when it comes to the maximum passenger count, this measure is not directly related to the amount of flexiblity or lack thereof caused by the different θ values. We believe this warrants future research which takes a closer look at how these discounts affect agents who travel specifically during the peak periods.

One thing that should be noted, is the high peak close to the end of the day in both Figures 4.5a and 4.5b. This is a clipping artifact and implies that a certain group of agents prefers to travel at the end of the day and suggests there is a problem with the calibration of these agents. Although the problem decreases when we increase θ, the problem does not disappear entirely, even when we have $\theta = \infty$. We ignore this problem for the time being, but it suggests that we should be careful in drawing conclusions based on these results, and it is an issue that should be addressed in the future.

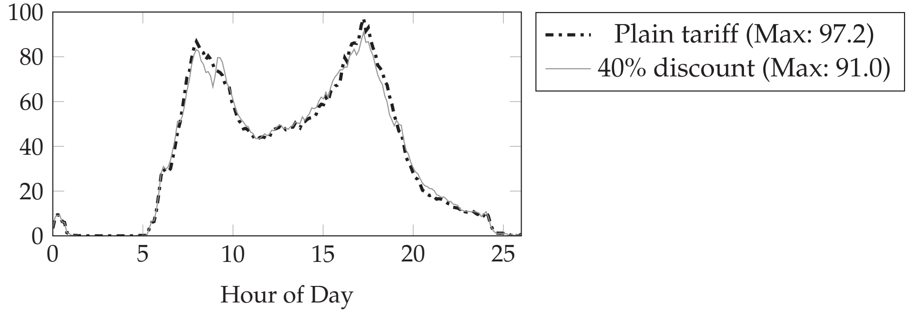
4.7 Discussion

Our results show that our proposed method of generating an agent population from a smart card dataset and perform a microscopic simulation where each customer is presented by an agent is achievable within a reasonable amount of time. Generating the agent population and performing a single run of the simulation (given that all routes are calculated) both take under two minutes of time.

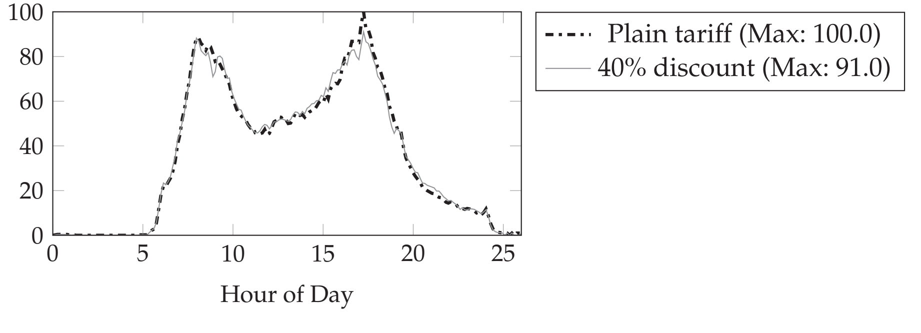
In the remainder of this section, we discuss different topics for future research. In Section 4.7.1 we discuss how to improve the demand generation itself. We discuss the possibilities with regard to calibration of the parameters used by the simulation in



(a) $\theta = 80$



(b) $\theta = 120$



(c) $\theta = \infty$

Figure 4.5: Vehicle loadings after 180 iterations for the minimum number of journeys before a smart card is used to create pattern based demand $\theta \in \{80, 120, \infty\}$. On the horizontal axis, the time of day is displayed. On the vertical axis, a normalized passenger count is displayed. The 1% and 5% discounts are not shown as they would be hard to distinguish from the lines currently visible.

Table 4.2: Normalized maximum passenger count and normalized revenue for different off-peak discounts and demand scenario's with $\theta \in \{80, 120, \infty\}$. The change resulting due to the discounts compared to the plain tariff within each scenario is also reported as a percentage point (pp). For the maximum passenger count, this change is also divided by the discount percentage, which can be interpreted as the pp change of the maximum passenger count per percent of discount.

	Max. # Passengers			Revenue	
Experiment	Level	Change (pp)	Change (pp) per discount %	Level	Change (pp)
Scenario $\theta = 80$					
Plain tariff	98.0			98.6	
1% discount	97.8	-0.2	-0.20	98.0	-0.6
5% discount	96.8	-1.2	-0.24	95.4	-3.2
40% discount	89.6	-8.4	-0.21	71.6	-27.0
Scenario $\theta = 120$					
Plain tariff	97.2			99.2	
1% discount	97.0	-0.2	-0.20	98.6	-0.6
5% discount	97.4	0.2	-0.04	96.2	-3.0
40% discount	91.0	-6.2	-0.16	72.2	-27.0
Scenario $\theta = \infty$					
Plain tariff	100.0			100.0	
1% discount	99.0	-1.0	-1.00	99.4	-0.6
5% discount	97.8	-2.2	-0.44	97.0	-3.0
40% discount	91.0	-9.0	-0.23	73.0	-27.0

Section 4.7.2. Section 4.7.3 addresses the question how we can incorporate additional datasets in order to distinguish different types of activities. Finally, Section 4.7.4 addresses the issue of validation.

4.7.1 Demand Generation

In our demand generation algorithm we make a couple of assumptions. The assumption that people who commute travel a lot, is very fundamental and probably realistic. The assumption that commuters have a fixed home and fixed working location probably often holds, but may be relaxed a bit: It can be broken by people who have more than one place to spend the night, or who have a job that has different locations that get visited in a regular patterns. With enough observations, it may be possible to detect such patterns as well. The assumption that people spend more time at home probably holds often as well, but we must be careful with regard to outliers: It may be possible that somebody switches mode while at work (either by taking a bike or a car). In such an event, it would be possible that our approach will reveal that somebody stayed for days at his working station, while this was not true in reality. However, the assumption that the variation in travel behavior of a passenger reflects his flexibility with regard to travel time is the most doubtful. Studies with more information regarding this assumption would be extremely valuable in improving our demand generation process.

While we have shown that our approach can efficiently generate an agent population from a real life smart card dataset, the fact that we have taken an approach that is very efficient and straightforward to implement has the disadvantage of being relatively crude. One may argue that we can introduce sophisticated pattern recognition and data-mining techniques in this process, in order to generate an agent population that is closer to reality.

4.7.2 Calibration

In order to reflect real life behavior more closely, calibration of our simulation is a required to use it in a decision support setting. The single global price elasticity for all agents is something that could be implemented on an individual level, without altering the structure of the utility function itself. One way to implement this, without modifying the general scoring function is to adapt our fare-module to vary the price based on the price sensitivity of an agent. We can use a personal transformation function for each agent that scales the fares down for insensitive agents and scales the fares up for sensitive agents. Another kind of sensitivity that could be valuable to include in the simulation, is the sensitivity to the crowdedness of vehicles. If vehicles become too crowded, additional delay can induce delays in the public transportation system. This aspect was mostly ignored in our current simulation.

The right values for the price elasticities will be very difficult to estimate from only check-ins and check-outs. The main problem in this regard is the fact that we do not know what possible alternatives were available and have been considered by the passenger, before he made his journey. In the field of discrete choice modeling, this kind of data is referred to as *revealed choice* data. In situations where surveys are conducted and the subjects are exposed to multiple alternatives from which they must select a single option, we get *stated choice* data. Within the field of discrete choice modeling, most of the research effort has been performed on analyzing stated choice data. Stated choice experiments are designed in such a way that properties such as the price elasticities of (groups within) the population can be esimated effectively.

In our case, it would be necessary to combine information obtained from stated choice experiments to calibrate the simulation obtained from revealed choice data. Some literature on how to combine revealed and stated choice data has been published during the 1990's of which Ben-Akiva et al. (1994) and Adamowicz et al. (1994) are two examples. However, Pelletier et al. (2011) mentions that there is little research in this area concerning smart card data.

4.7.3 Extensions based on additional datasets

One possible way to move our pattern based demand closer to real activity based demand models is by combining the smart card dataset with other datasets. A promising approach might be to look at regional information of stations. We could use such datasets to construct profiles of stations, that allow us to make better guesses with regard to the activities that can be performed around the stations. If a station is close to a large industrial plant or office buildings, it is very probable that passengers traveling there do so because they have to work. A station close to a shopping mall will not only attract the employees of the shops, but customers as well. Local stations that coincide with a railway station or an airport are likely to attract passengers that want to travel further, or want to travel home. Stations in residential areas will likely serve as a home station, or as a station that gets visited by passengers who want to visit friends or family. We propose to use data provided by the OpenStreetMap project (OpenStreetMap Foundation, 2012), since this contains tags that contains information on available activities at certain locations.

After we generate profiles for all our stations with such information, we can take this information into account while recognizing patterns. This would allow us to make better guesses of the temporal flexibility of passengers for which we do not have a large enough set of journeys. Suppose we observe a passenger who starts his day with a journey from a residential area to an area with a lot of office buildings and stays there for six hours. He travels to an area with a shopping mall and stays there for one hour, after which he travels home. Even if we never observed any other journeys by this passenger, we can still make an educated guess about what he was doing and thus to what extent he could have been flexible. However, this calls for much more sophisticated statistical techniques than the one we are currently using. Depending on the kind of questions we want to study, it may or may not be worth the effort to go this far.

4.7.4 Validation

Validating a simulation like this is not a trivial task. One aspect that we can validate is the question whether the simulation can be used as a predictive tool for the movement of passengers through a public transport network. The straightforward way to do this is by splitting the dataset at a certain moment in time. We then use the first part of the dataset to generate agent populations and compare the outcomes to what is observed in the second part of the dataset. At first, we should choose a moment within a period where no policy and scheduling changes have occurred. If we can pass this test, we can raise the bar by choosing the moments at which a policy change has occurred, such as the introduction of a new schedule or new pricing schemes.

Another aspect that we may want to validate, is the question whether the emerging activity patterns of the agents represent the real-life activity patterns of the passengers represented by the agents. Validating this aspect requires much greater effort than validating the movements of passengers. One approach could be to use survey data containing activity logs registered in diaries and compare the diaries to the activity plans in the simulation. There may be some privacy issues with this approach, since it would require that we link the smart card id's to the participants, in order to match a diary to an agent. A possible workaround is to generate dummy check-in/check-out data from the diaries by generating a check-in and a check-out for the journeys documented in the diaries. We could then use this dataset as if it were a smart card dataset and investigate to what extent the generated activity patterns of the agents reproduce the original activity plans.

In a similar way, we can consider the study of other location tracking datasets, such as triangulation logs from mobile phone operators or the location logs from the mobile phones themselves. The main advantage is that such a dataset contains more details on the whereabouts of individuals, which gives more opportunity to estimate what they are doing. When using smart card data, we may observe that a person checks out at a station near a shopping mall and checks in four hours later. However, we have no data to decide whether it is probable that this person has been shopping or that this person has been working as an employee at one of the stores. If we have a mobile phone log, we may observe that the person has visited a great number of stores during these four hours. This would be evidence that he was not working as an employee. However, mobile phone data often comes as a set of observations in time and space without device id's to protect the privacy of their owners. As a result, this direction of research has severe challenges.

4.7.5 Economic Modelling

In practice, revenue management studies are often driven by the price elasticity of the demand, which models the relation between the change in demand and the price of travel. In an activity based simulation model these elasticities are an *emergent* property of the scoring function that uses the activities and travelling to compute utilities. As a result, activity based models depend on parameters for every activity and parameters for travelling. Based on all these parameters, the price elasticity of the demand becomes an emergent property rather than something which is put directly in the model. One potential advtange of this approach is that in some situations it can be easier to model activities of the population that to perform economic experiments to obtain the price elasticity. Another potential advantage is that it may be easier to include non-linear interactions between the demand and congestion, other modes of transport or properties of the activities than to model the price elasticity for each and every situation that may occur.

There are also disadvantages to this approach. First, the behavior of the population is quite sensitive to the parameters of the activity based model. In one preliminary study we increase the typical duration of the home activity to twenty hours, which resulted in a time shift of the afternoon peak. This raises the important question how the different parameters of the activity based model influence the price elasticity of the demand, and to which extent this relation is realistic for practical applications.

An interesting area for future research is to explore in detail how the price elasticity is influenced by the different parameters of the model and the price elasticity of the demand. To gain better understanding of the emergence of a macroscopic such as the price elasticity, it makes sense to use simple scenarios with a single public transport line with a few stops and investigate the trade offs a single agent can make with regard to travel times and the resulting utilities. While there might be potential for analytical results in this area, it is also possible to simulate a number of scenarios with different parameters and compare the plans and corresponding utilities emperically.

4.8 Conclusions

We have shown how we can use smart card data to generate different types of demand. We developed an agent-based simulation that allows us to analyze the movements of the agents through our multimodal public transport network. We experimented with different settings for the number of trip-based agents and with three discount policies in the off-peak hours. Finally, we discussed several opportunities for future research.

As soon as we sorted our dataset in such a way that we could process all journeys customer by customer in chronological order, demand generation could be done very efficiently. We used simple rules to determine whether a customer should be modeled using *trip* based, *tour* based or *pattern* based demand. We have evaluated the impact of different thresholds for the *pattern* based customers on the resulting approximate equilibrium. We have also seen that an off-peak discount can be used to let a part of the agent population shift their travel times. In our case, this lead to a lower revenue. However, the effect on the required capacity must be taken into account when making a trade off between costs and revenue.

There are many opportunities for future research. First, our simulation can greatly benefit from proper calibration. Configuring the behavior and economic parameters of the agents is far from trivial. This requires surveys with representative groups and expert knowledge that the various public transport companies might or might not be willing to share. The amount of work this requires will likely result in a model that is better suited for practical applications, but it does not necessarily answer how we can use methods and ideas from *complexity science* in the modeling and optimization of public transport processes. Including heterogeneity in the price sensitivity of is also an important step to include more accurate economic behavior of the agents, although this would be mostly a software development challenge that should not be too difficult to overcome, given the open structure of the MATSim framework.

Other opportunities for future research are the method for demand generation, which can be improved upon. One way is to take a closer look at the smart card data itself and apply data mining techniques to create a number of different behavioral types. Another opportunity is to combine the smart card data with additional datasets. We believe this line of research to be better suited within the framework of *complexity science*, as it is concerned with the relationship between emerging patterns and microscopic travel behavior.

We believe that an improved version of our simulation, where economic behavior is realistic and properly validated, can be helpful in both the design of revenue management systems, including location based and modality based tariff schemes and other fields of study within a public transport context.

5

Detecting Activity Patterns from Smart Card Data

Co-authors : Evelien van der Hurk, Leo Kroon, Ting Li and Peter Vervest
This paper has been accepted after peer review for presentation at, and publication in the proceedings of, the 25th Benelux Conference on Artificial Intelligence (BNAIC2012).

5.1 Introduction

One of the most valuable pieces of information during the development and operational planning of passenger transportation systems is passenger demand. Understanding how demand develops allows governments and public transport operators to assess the profitability of infrastructure investments. By changing the infrastructure or by developing a new area, new travel routes and new purposes to travel are created. Transportation is a major concern in the cost-benefit analysis of such large projects. Passenger demand also serves as input for the design of public transport systems, where they are important when decisions on which lines to operate, the service frequencies and the assignment of vehicles of different capacities to each service are made.

Traditional demand models typically estimate an origin-destination matrix of trips and use a traffic assignment model to map routes to OD-pairs in the transportation network. Typical data sources for such matrices are counts of observed travellers. Especially in public transport counts can be used to obtain an origin-destination matrix within a reasonable processing effort. However, counts cannot be obtained for cases where we consider a change in infrastructure or service. To overcome this problem other data sources, such as data related to land use and travel-choice surveys are used to make predictions of demand in other scenarios. One of the drawbacks of using origin-destination matrices is that it is not straightforward to make a matrix time-dependent, analyze the behavior of specific groups of travellers (outside of an origin-destination gorup) or to model more detailed interactions between passengers and public transport, such as for example due to crowding.

Activity-based models (Axhausen and Gärling, 1992) provide an improvement in this regard. The main idea of this paradigm is that transport demand emerges from many individual desires to perform certain activities at different locations at certain times. Instead of defining a large matrix as demand, a list of activity plans is created. In a simulation agents can perform these activities in a simulated transportation network. It is possible to observe how the agents decide to travel from activity to activity within the simulation. An example implementation of such a model is the open source agent-based transport simulation package MATSim (Horni et al., 2016), that has been applied at different locations around the world. The input required for such models consists of individual day plans that define a chain of activities. Since this input data cannot be directly deducted from an OD-matrix, random plans generated from economic and geographical data are often combined with travel diaries collected through surveys.

In this chapter we develop a method to deduce and analyse activity patterns and activity sequence patterns within the time dimension. We define an activity as a combination of a time interval and a location. These activities are reconstructed from the set of trips stored in the data for a specific person. Using both clustering and labeling methods, we identify important activity time intervals and analyse common activity chains. We consider a time interval to be important if it represents at least 10% of the activities at a station in the network. We are not only able to identify home-work patterns, but also identify shorter activities. Moreover, the activity chains provide information on the order of different activities. We aim to extend our method to include spatial dimensions in the future, by labeling stations into groups based on the temporal patterns outputted. We believe that the results obtained using our method can provide public transport operators insight into how their network is being used and give valuable input for activity based models.

5.1.1 Related Work

Pelletier et al. (2011) present an excellent general review of smart card data research in public transport during the years 2000-2010. As this is an extensive literature review, we only present a short overview of research focused on activity analysis based on smart card data.

Agard et al. (2006) analyse a binary vector indicating smart card activity during four fixed time slots, defined by the public transport operator. They find four main travel patterns using hierarchical clustering, the top two of which correspond to a home-work-home pattern and a home-study-home pattern. Morency et al. (2007) focus on the variation in temporal patterns using a k-means clustering algorithm. They also consider vectors of 24 binary values indicating whether a passenger has boarded a vehicle during the corresponding hour of the day. Using clustering with the Hamming distance measure and the component wise median to derive cluster centroids, they are able to derive regularity indicators from the raw data. Devillaine et al. (2012) present an analysis focused on the temporal distribution of activities based on smart card data from both Santiago, Chile and Gautineau, Canada. Their classification is based on both temporal aspects as well as card type. The assigned classes are work, study, home and other. They find that the temporal distribution of activities in Santiago differs from Gautineau. Activities classified as other have peaks at their starting times when they start more often around noon or four in the afternoon in the Gautineau network, while they are more evenly distributed in the Santiago network.

A different methodology to analyse spatio-temporal patterns is to calculate eigenbehaviors (Eagle and Pentland, 2009). The general idea of the method is to apply Principal Components Analysis on vectors of binary variables representing time slot/location combinations. While this method is usually able to reduce a matrix of vectors to a few dominant eigenvectors, the fractional nature of the eigenvectors makes them complicated to interpret, especially if the goal is to create input for activity based models.

5.2 Journeys and Activities

The central idea of *activity based modelling* is that mobility patterns of individuals can be regarded as a sequence which alternates between *journeys* and *activities*. For example, we can have a sequence such as

$$(t_1, l_1) \cdots (t_2, l_1) \to (t_3, l_2) \cdots (t_4, l_2) \to (t_5, l_3) \cdots (t_6, l_3) \to \ldots$$

where the variables t indicate time instances such that $t_i \leqslant t_{i+1}$ and the variables l indicate locations. It has been proposed numerous times that a good model of activities performed by individuals will generate high quality travel demand, i.e. the $\to$ arrows representing journeys can be inferred from the dots representing activities. However, many data sources collected by and available for transportation agencies contain either observations at a short time instance (such as mobile phone data) or information on journeys (such as smart phone data or number plate tracking at entrances and exits of highways).

As operators and individual passengers adopt information technology at an increasing rate, we now have numerous data with information on the locations of individuals. For the purpose of demand modelling, we have a number of guiding principles for an analysis system:

Scalable In case the number of observations for individuals is too large to store on one computer, it would be useful if we can still compute relevant statistics. We will assume that the data for a specific individual or a specific location is available at a single node of computation, and that the number of locations under consideration is very small compared to the number of observations.

Anonymous Aggregate Data Since individual travel patterns contain a lot of privacy sensitive information, we prefer that the system is able to collect interesting aggregate data without the need to keep very long histories of individual observations. In other words, the aggregate data should be anonymous.

5.3 Smart Card Data

The Dutch smart card system, called "OV-Chipkaart" is a nation-wide smart card for payment of public transport journeys across modes and operators. The system is operated by the common smart card authority "Trans Link Systems", which collects the transactions and provides the operators with the data of their customers. This is raw transactional data where each record contains at least a unique media ID of the smart card, date and time of the transaction, an ID specifying the station or stop where the transaction took place and the type of the transaction (i.e. check-in or check-out). One of the important features of the Dutch implementation is that passengers both need to check in when they start their journey and check out at the end. As a result, we do not need to estimate alighting points.

In order to analyse the intervals corresponding to activities in the network, we have extended our implementation discussed in Chapter 4 in order to extract the activities from the raw smart card data.

5.3.1 From raw transactions to journeys

The first step in preparing our data for analysis is to derive a data set of journeys from the raw smart card data. In order to do this, we sort the raw smart card data based on the media ID and the time stamps, such that we can easily process consecutive transactions card by card. Every time we detect a check-in followed by a check-out while passing through the data, we generate a trip containing a departure time, departure location, arrival time and arrival location. For some modes (bus and tram) a journey may consist of several trips. After the construction of trips, we merge all trips that take place within the operator specified allowed transfer time. If we end up with journeys that start and end at the same station, we remove them from the final set of journeys.

5.3.2 From journeys to activities and time intervals

After we have obtained a sequence of journeys $j_1, j_2, \ldots, j_n$ for each smart card, ordered by time, we continue with the detection of activities. If journey j_i's arrival location is equal to journey j_{i+1}'s departure location, we create an activity at the common location from the arrival time of journey j_i until the departure time of journey j_{i+1}.

As it is possible that activities span multiple days, we simplify them by projecting them onto a modular ring. Let us first pick a number of time slots U. Throughout this chapter we will work with hourly time slots, so $U = 24$. All calculations involving the intervals will now be done on the modular ring $\mathbb{Z}_U$. Under the assumption that $\mathbb{Z}_U$ represents a day, the begin time of the activity is projected onto the ring, rounding the final time slot down after scaling, while the end time of the activity is rounded up after scaling. Given an interval x, the start and end times are denoted by the pair (x_b, x_e), which starts at a time slot $x_b \in \{0, 1, \ldots, U - 1\}$ and ends at time slot $x_e \in \{0, 1, \ldots, U - 1\}$. As a result, a time slot can be an "overnight" time slot in case $x_b > x_e$. For such time slots, it is not correct to take the difference $x_b - x_e$ to calculate the duration of the time slot, as time moves forward. To overcome this fact we define the duration x_d of an interval x as follows:

$$x_d = \begin{cases} x_e - x_b & \text{if } x_e \geqslant x_b \\ x_e + U - x_b & \text{otherwise} \end{cases}$$

Note that this formula does not make a distinction between activities that are longer by a multiple of U, thus activities which have an actual duration of 32 or 80 will both have an x_d of 8. We do this because it simplifies the analysis and presentation of the data a bit, but for practical applications it should be considered to take information about activities that take multiple days into account.

5.4 Extracting Frequent Time Intervals by Clustering

As the number of different intervals observed at each station is likely to be too large for regular interpretation, we will apply a clustering algorithm in order to obtain a compact description of the types of intervals observed at the station. As the dissimilarity of two time intervals may depend on the context of the activities, we introduce a parameterised penalty function that will be used to as a distance measure for clustering. As an example of such differences, consider that activities at an office will likely have high similarity in the starting time of the activity, while shopping or entertainment activities are more likely to have similarity in the duration.

After processing the raw smart card data, we end up with a set of stations S and a multiset I_s of observed intervals at a each station $s \in S$. We then apply the k-means++ algorithm (Arthur and Vassilvitskii, 2007) discussed in Appendix A on each multiset I_s. The advantage of the k-means++ over the traditional k-means algorithm is that it is $O(\log k)$ competitive due to a sampling method for the initial clustering that favours centroids that are far away from each other. Since there are many stations in the network, we also propose a method to aggregate the cluster outputs to a full network level. The reason we do not apply the clustering algorithm on the union of all I_s multisets is that we are also interested in time intervals that occur frequently at a station that does not serve a large part of the total demand.

Finally, as the results of the clustering algorithm may vary with the random initial configuration, the parametrization of the penalty function and the choice for k, we run our clustering and aggregation method multiple times in order to get a feeling for the robustness of the cluster centroids. As there there is a finite discrete set of possible centroids, we measure the frequency in which each centroid occurs in the output of the different runs of the algorithm.

5.4.1 The parameterised penalty function

In order to assign different penalties to differences between start time, duration and end time of the activities, we introduce a vector $\theta = (\theta_1, \theta_2, \theta_3)$. Here, θ_1 and θ_2 control the penalties if either the duration, start time or end time is equal, while θ_3 controls the penalty if these values are all different. Our penalty function that represents the difference between two intervals x and y is defined as follows:

$$d_\theta(x, y) = \begin{cases} \theta_1(x_d - y_d)^2 & \text{if } x_b = y_b \vee x_e = y_e \\ \theta_2(x_b - y_b)^2 & \text{if } x_d = y_d \\ \theta_3(|x_b - y_b| + |x_d - y_d|)^2 & \text{otherwise} \end{cases}$$

Note that our penalty function d is heuristic in nature, as it can end up awarding a large penalty to a pair of activities for which the end times are close to each other, but a much smaller penalty to a pair of activities for which the start times are as close to each other. As one of the aims of our method is to aggregate many different clustering outcomes to decide which activity patterns are common in the data set, the focus of our research was not to make this heuristic penalty function mathematically sound. The design of a mathematically sound distance measure for our methodology is a promising direction for future research.

In addition to the distance measure, we also need a way to calculate the centroid of a cluster. Since we work with $\mathbb{Z}_U$, the set of all intervals is given by $\mathbb{Z}_U^2$. In case of $U = 24$ this gives us 576 intervals. As a result, the best cluster center within a cluster of size n can be brute forced in $576 \cdot n$ calls to the distance measure.

5.4.2 Calculating the relevant cluster centroids and their robustness

In order to aggregate the clustering output of the individual stations, we decided to work with a threshold-based rule. This rule check for a given threshold t whether an interval (x_b, x_e) is relevant if there exists a station $s \in S$ such that (x_b, x_e) occurs as a centroid in the cluster output of the multiset I_s and that cluster contains more than $t|I_s|$ elements of I_s. The set of all intervals that adhere to this criterion can be calculated using Algorithm 5. We also keep track of a weight map w, which registers

Algorithm 4: Iterative loop used to calculate the robustness fraction.

Input : A set $\mathcal{C}$ of configuration parameters for runExperiment, a cutoff number m

Output: A table with for each (x_b, x_e) interval the robustness fraction

```
Method calcRobustness(𝒞) :
    r ← new table of dimension U × U filled with 0-values;
    foreach (θ, k, σ, t) ∈ 𝒞 do
        (J, w) ←runExperiment(θ, k, σ, t);
        foreach (x_b, x_e) ∈ J do
            if w((x_b, x_e)) ⩾ the mth highest value in {w(x) : x ∈ J} then
                r[x_b][x_e] ← r[x_b][x_e] + 100/|𝒞|;
            end
        end
    end
    return r
end
```

the fraction of the population covered by the interval in the cases where it exceeds the threshold.

Since the output of the clustering algorithm, and therefore the output of the *runExperiment* method can vary for different configurations of the parameters, we decided to apply it multiple times, keeping track of how often each interval shows up. Since the number of intervals in a single result set can still be quite large, we truncate the result set to the m highest scoring intervals according to the weight map w. We then count the number of times an interval is in a truncated result set and report this as the fraction of the total number of experiments as the "*robustness fraction*". The loop we use to calculate the *robustness fractions* is presented in Algorithm 4.

5.5 Labeling and Activity Chain Analysis

After we have applied the clustering algorithm to learn important time intervals in the network, we want to learn something about the relationship between the activities that take place during these intervals. Utilising the output of the clustering algorithm, we can propose a labeling algorithm that assigns a label to each interval. This algorithm then allows us to transform the chains of activities observed in the data of the separate passengers into chains of activity labels.

Algorithm 5: Iterative loop used to calculate relevant time intervals in the network

Input : Distance measure parameters θ, the number of clusters k, a random seed σ, a threshold t

Output: A set of relevant intervals R, a weight map w

```
Method runExperiment(I, θ, k, σ, t) :
    foreach station s ∈ S do
        R ← ∅;
        C_1, ..., C_k ← k-means++ applied on I_s with penalty function d_θ and
          random seed σ;
        for i ∈ 1, ..., k do
            if |C_i| / |I_s| ⩾ t then
                x ← centroid of C_i;
                R ← R ∪ {x};
                w(x) ← w(x) + |C_i| / (|I_s||S|);
            end
        end
    end
    return (R,w)
end
```

5.5.1 Developing the labeling algorithm

In Devillaine et al. (2012) it was observed that there are differences in the extent to which different time intervals show up in different public transport networks. Since our intervals are described by two time slots, it is straightforward to visualise the robustness fractions in a grid containing all possible intervals. Such a plot gives great insight in the extent to which time intervals are important. We can use the plot to construct a tree-based labeling rule by checking proposed rules against the robust intervals in the plot.

An important aspect to take into account during the development of the labeling rule is the interpretability of the chosen labels. As our focus is currently mostly exploratory, we decided to focus on labels that are easy to interpret, such as long, short, early, late and overnight.

5.5.2 Analysing consecutive activities

Utilising the labeling procedure developed in the previous section we can now analyse chains of activity labels. We count all consecutive pairs of activities that are performed by the same person and are connected by a single journey. The resulting table of counts can then be interpreted as the adjacency matrix of a weighted directed graph, where the nodes represent the activity labels and the arcs represent the "followed by" relationship as observed in the data. There are many software packages available that allow us to visualise and analyse such networks. During our analysis, we have worked with Gephi (Bastian et al., 2009).

We can count activity chains of an arbitrary length in a similar fashion. We believe that counting chains that are very long will not give a lot of insight, as passengers are not likely to perform many activities within a single day. However, smaller chain lengths, such as three or four activities, could be interesting as these patterns are likely to represent behaviour over one or two days. For this reason we added a triplet counting routine to our implementation of the processing algorithm for the adjacency matrix generation.

5.6 Experiments and Results

We have applied our clustering method on urban smart card data set from a Dutch network, containing four months of transactions. The data set contains roughly $22 \cdot 10^6$ journeys and $12 \cdot 10^6$ activities. We calculated the *robustness fractions* using the method described in Section 5.4.2. Our set of configurations contained all combinations of the following: For k one of $\{6, 8, 10, 20\}$, $\theta \in \{1, 2, 4\}^3$ with the constraint that $\theta_3 \geqslant \theta_1 \wedge \theta_3 \geqslant \theta_2$, one of two random seeds, $m = 40$ and $t = 0.1$. The total number of configurations is 112. The resulting table is visualised in Table 5.1.

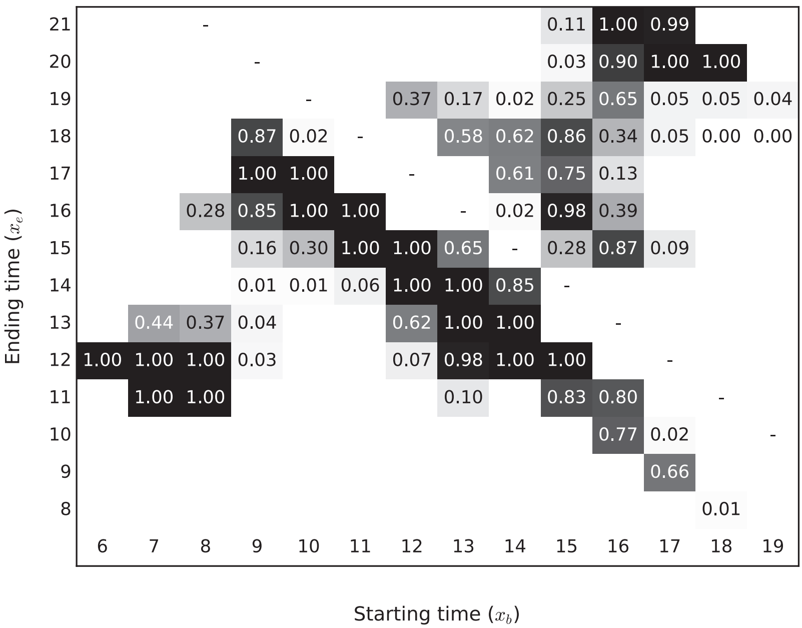
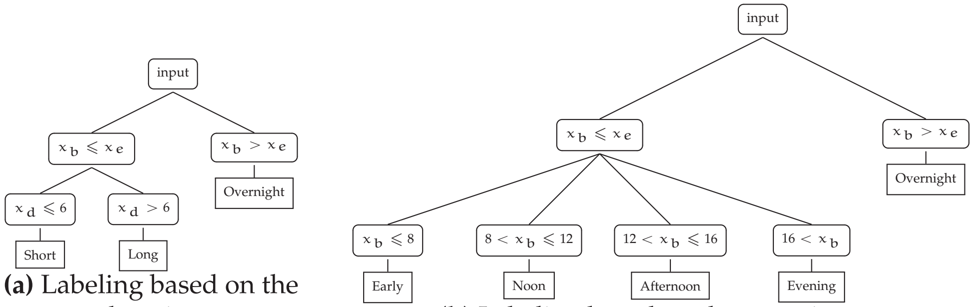


Figure 5.1: The table of the robustness fractions of the intervals as calculated by our clustering and aggregation algorithm. Every cell corresponds to an activity interval with the combination of a starting time and ending time. Activity intervals at empty cells are never detected for any configuration of the clustering method by the aggregation algorithm, while cells with a value of 1 are detected for all configurations of the associated activity interval.



(a) Labeling based on the duration

(b) Labeling based on the start time

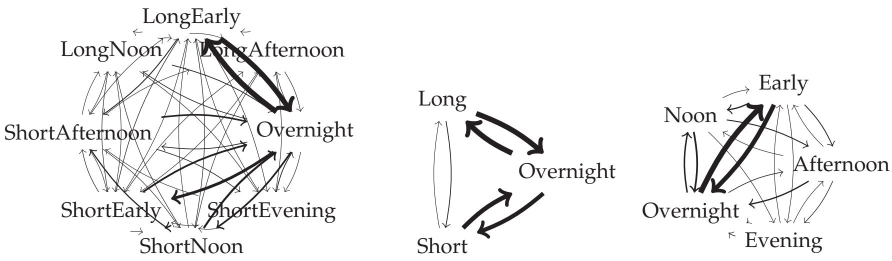
Figure 5.2: Labeling trees for the labels of an interval

Many of the highly robust intervals in Figure 5.1 are typically associated with commuting patterns. However, many shorter intervals that start after nie are quite robust as well. Additionally, intervals with a duration of precisely six or seven hours are very infrequent. This tells us six hours is a natural boundary to distinguish between short and long activities. The proposed labeling on the duration is presented in Figure 5.2a. Distinguishing between starting times appears to be more complicated. Before 9:00 short activities rarely begin, so 8:00 seems to be a good boundary for early activities. At 13:00 it seems that among the shorter activities, the intervals that are one hour longer than those starting before 13:00 become robust as well. As a result, we pick the second boundary at 12:00. Finally, after 16:00 the intervals or not very robust, so this gives us the final boundary. As it would be hard to interpret different labels for overnight activities, we introduce a single label "Overnight" for activities that have $x_b > x_e$. The resulting tree is presented in Figure 5.2b.

We visualised the trees for labeling on duration and start time based labeling separately, but they can be combined into one large tree. We can now apply the different trees to count the pairs of sequentially occurring activity labels. We visualised the pairs observed this way as a network in Figure 5.3

Some interesting patterns can be observed in Figure 5.3. First, the most prominent pairs of activities are those between an overnight interval and intervals that are early and long. These are typically time intervals associated with home-work and home-study patterns and are thus within expectation high ranking. A more interesting pattern occurs between overnight and noon activities. There is less interaction between early activities and noon activities than between overnight and noon activities.

Let us now consider the top ten of triplets occurring in the activity label chains. The most dominant triplets are typically associated with home - work - home like chains. The fourth and fifth triplet in the full labeling describe a single activity during noon. Here we might be a bit careful in labeling the overnight activity as home: Maybe some people travel to their work by car and use the public transport system during lunch



(a) Combined labeling (b) Duration based labeling (c) Start time based labeling

Figure 5.3: Network visualisation of the adjacency matrices based on the different labelings.

time to visit a nearby location. The tenth triplet shows a pattern where two activities are started within the noon window. There is also evidence of people performing a long activity one day and a short activity the next day, and vice versa, as witnessed by triplets eight and nine in the duration based labeling.

When we compare these results to the analysis of the "other" activity label considered during the analysis of Gautineau data by Devillaine et al. (2012), we see that the third triplet in the start time based labeling suggest a possible peak around 12:00 for at least some of the stations. However, they also found a peak around 16:00, which would be the afternoon label in our case. However, if we consider labels in the start time table, only the sixth and ninth triplets represent evening activities and both are not as strong as the single noon triplet.

5.7 Conclusions and Future Work

We have developed an approach to cluster temporal intervals derived from activity data at a station level using a parameterised penalty function nd to aggregate the results, such that we obtain the most interesting time intervals in the data. We repeat this process to obtain robustness fractions. Based on the robustness fractions, we constructed a tree-based labeling procedure. The labels allow us to find the most frequent pairs and triplets of activity types observed in individual activity chains. While the typical intervals associated with home and work activities are dominant, we are able to identify shorter activities as well and provide some insight on their relation to other activities within the activity chains of individual passengers.

Our current approach still has some drawbacks. The modular ring $\mathbb{Z}_U$ with $U = 24$ is a quite rigorous simplification, as we cannot distinguish between an activity that takes one hour and an activity that takes 25 hours. While this simplification allows

Label 1	Label 2	Label 3	Percentage of observed triplets
Full labeling			
Overnight	LongEarly	Overnight	19%
LongEarly	Overnight	LongEarly	16%
Overnight	LongNoon	Overnight	4%
Overnight	ShortNoon	Overnight	3%
ShortNoon	Overnight	ShortNoon	2%
ShortAfternoon	Overnight	ShortNoon	2%
Overnight	ShortEarly	Overnight	2%
LongNoon	Overnight	LongNoon	2%
ShortAfternoon	ShortAfternoon	Overnight	2%
Overnight	ShortNoon	ShortNoon	2%
Duration based labeling			
Overnight	Long	Overnight	23%
Long	Overnight	Long	20%
Short	Overnight	Short	10%
Short	Short	Short	9%
Overnight	Short	Overnight	7%
Short	Short	Overnight	6%
Overnight	Short	Short	6%
Short	Overnight	Long	6%
Long	Overnight	Short	4%
Overnight	Long	Short	2%
Start time based labeling			
Overnight	Early	Overnight	20%
Early	Overnight	Early	19%
Overnight	Noon	Overnight	7%
Noon	Overnight	Noon	5%
Noon	Overnight	Early	4%
Afternoon	Overnight	Noon	3%
Early	Overnight	Noon	2%
Afternoon	Overnight	Early	2%
Overnight	Afternoon	Overnight	2%
Overnight	Early	Noon	2%

Table 5.1: The most frequent triplets of consecutive labels for each labeling method and the percentage with which they occurs among all triplets detected when applying that particular labelling method. Triplets are generated based on the consecutive activities of individual smart cards. If the frequency of a triplet is 20% this means that if you look at three consecutive activities of a single smart card, in 20% of the cases these labels correspond to the labels of the triple.

us to get a general idea of what is happening within the system without having to look at too many numbers, it is likely more caution is necessary if we want to construct the input for activity based models. Moreover, we ignore the distinction between weekdays and weekends, which has a very significant impact on travel behaviour. For the implementation of a valid simulation, it will be necessary to make this distinction. If we introduce these detailed descriptions of the activity intervals, it would also be necessary to reconsider the distance measure used. In future work a more rigorous mathematical basis could be used in its design as part of this process. Finally, we constructed our labeling algorithm by hand. A question is whether we can use automatic classification algorithms instead of our manually constructed labeling procedures.

Aside from further refinements of our methods, such as reducing the number of parameters to set and varying distance measures, there are two main topics for future research. First, we can use either the clustering output at the station level or the complete distribution of intervals observed at the stations to identify similar classes of stations. If we are able to reduce our stations to a small number of important classes, we can attempt to include some spatial aspects in the analysis of the activity chains. Finally, it can be consider how our findings can be used to improve demand generation procedures for agent-based simulation studies similar to the study in Chapter 4.

6 Exploratory Analysis of Time-Space patterns in Smart Card Data

Co-authors : Evelien van der Hurk, Leo Kroon and Peter Vervest
This chapter is a working paper.

6.1 Introduction

The introduction of smart card ticketing systems has resulted in a wealth of data for public transportation operators compared to analogue ticketing systems. Previously a lot of effort was required to make crude estimates of passenger volumes, often split in peak and off-peak volumes. Smart card ticketing provides detailed information related to time and location of arrivals and possibly departures for each passenger. It may come as no surprise that some practitioners think that with the availability of smart card data, the operators now know *everything* there is to know about the passengers.

Unfortunately smart card data does not contain all information relevant to transport planning. Researchers in transport demand modeling (Ben-Akiva et al., 1994) are very aware of the difference between *stated choice* data, i.e. information that states which travel option a passenger considers as best from a set of given alternatives, and *revealed choice* data, i.e. information about which actual choice a passenger took but not the considered alternative travel options.

Parallel to the introduction of smart cards in public transport, disruptive new business models were introduced in *private transportation* often driven by informative smart phone applications that give detailed expectations about the journey a prospective customer is considering. Such developments give travelers more freedom in planning their journeys. In the nearby future it may also become possible that fleets of on-demand automated vehicles are introduced, decreasing the need for car ownership in favor of on-demand automated vehicles. An advantage of such a system is that, if designed properly, vehicle utilization can be increased as currently most cars are not utilized for the majority of the time. The big challenge for this business model is whether it can be made reliable. If the chances of getting an on demand vehicle are very low during peak-hours, there is a still an important reason for many individuals to keep owning cars. In order to decrease car ownership, it is important that the information technology used to organize such systems is user friendly and reliable. It is likely that this will require models that go beyond the scheduling of individual trips, but take activity patterns of the customer base into account.

In order to provide good service qualities to the passengers of public transport and other systems where vehicle sharing is employed to increase utilization, it is important to understand what goals passengers are trying to accomplish with their journeys and what alternatives they consider as acceptable for achieving these goals. A passenger who has to catch a plane at the airport has other requirements related to the journey than a passenger who wants to have coffee with family. Nowadays, this type of information is barely taken into account as public transport operators focus on moving certain volumes of peoples between origin destination pairs as efficient as possible. With the advent of potentially cheaper and smarter private transportation systems on the horizon, there is a great need to design public transport systems and the related information systems from the perspective of the passenger.

In previous chapters we have considered home-work patterns in smart card data as well as temporal patterns. In this chapter we focus on patterns that look at temporal and spatial properties, as activities performed by individuals happen at a certain time interval at a certain location. In this chapter we consider how we can improve upon these methods by proposing a way in which the generated patterns can be validated. As we do not know the underlying pattern for actual smart card data, we develop a way to produce synthetic smart card data for which these patterns are known. We generate data for a mockup scenario based on the city of Utrecht in the Netherlands. We evaluate a clustering approach that groups stations based on similar temporal profiles and show that this method can also be applied to other methods that extract behavioral patterns from smart card data.

The remainder of this chapter is organized as follows: In Section 6.2 we consider and formalize the general structure of smart card data and its relation to activities. In Section 6.3, we introduce an algorithm that can generate synthetic smart card data based on predefined activity patterns that can also be used in the validation step. In Section 6.4, we propose a way to make groups of stations based on the distributions of labelled time intervals at each station. In Section 6.5, we introduce the demand patterns and synthetic smart card data set that we use in our experiments. In Section 6.6, we compute groups of similar stations based on the synthetic data set and validate how well these groups of stations match with the predefined demand patterns. In Section 6.7, we conclude with a brief discussion of the obtained results and sketch a direction for future research on the topic of methodologies for smart card data analysis.

Table 6.1: Example dataset of processed smart card data

Smart Card ID	Departure Timestamp	Arrival Timestamp	Departure Location	Arrival Location
2348	2016-03-14 9:06	2016-03-14 9:18	Central Station	Main Square
2348	2016-03-14 15:03	2016-03-14 15:21	Main Square	Central Station
2348	2016-04-14 11:01	2016-04-14 11:03	Central Station	Shopping Mall
5231	2016-03-14 10:03	2016-03-14 10:33	Suburban Street	Main Square
5231	2016-03-14 14:37	2016-03-14 15:07	Main Square	Suburban Street
...	...	...	...	...

6.2 Activities and Smart Card Data

Smart card data come in various formats and there are variations in the technical details of the systems that collect the data between diffirent public transport operators. Some ticketing systems collect only check-ins, while other systems collect both check-ins and check-outs. This latter system reduces the need to estimate the check-out locations afterward, which is an advantage for the analysis. However, a number of methodologies that estimate check-out locations based solely on the check-ins are discussed in Pelletier et al. (2011).

We assume a smart card database consists of records that contain a smart card id (often the serial number of the media used), a time stamp, the location of the transaction, the type of action (check-in or check-out) and possible additional information such financial data (which we neglect in this chapter).

Often it is easier to link a check-in and a check-out to create a table of *journeys* instead of raw transactions. For this purpose it is a matter of sorting the raw transactions based on the smart card id and on the time stamp. Consecutive check-in and check-out pairs can be easily combined this way, yielding a dataset with the structure displayed in Table 6.1.

From this dataset of journeys, we derive a similar dataset of activities. In order to describe this process, we first formalize the symbols used to describe both the journeys and the activities.

6.2.1 Formal definitions of journeys and activities

Let us now introduce the formal definitions that make up journeys and activities. We work with the assumption that individuals alternate between traveling (journeys) and performing activities. This can be done without loss of generality, if we allow activities to have no duration (i.e. their start and end times coincide) and journeys that have the same location as departure and arrival locations. To define activities, we first define how to represent time and space:

- Time instants are modeled by choosing a temporal unit (e.g. minutes or seconds) and a time offset (for Unix timestamps this is the start of the first of January 1970). As a result, we can use common arithmetic on the natural numbers to work with time instants.

- Time durations are also modeled using the natural numbers, with the same temporal unit as we are using for time instants.

- A set of locations L that occur within our dataset.

Using our chosen representation of time and space, we can now define how to model activities:

Definition 2. *An activity A_i is a 3-tuple (a_i, d_i, l_i) where $a_i \in \mathbb{N}$ is the time at which the activity starts, $d_i \in \mathbb{N}$ is the time at which the activity ends and a location $l_i \in L$ where the activity takes place. The duration of an activity is computed by $d_i - a_i$.*

Suppose that some individual performs n activities at different locations. We can then consider following activity sequence for this individual:

$$A_1 = (a_1, d_1, l_1), A_2 = (a_2, d_2, l_2), \ldots, A_n = (a_n, \infty, l_n)$$

In this sequence, it must hold that $a_i \leqslant d_i \leqslant a_{i+1}$ for all i, since time always goes forward. Associated with a sequence of activities, we have a corresponding sequence of *journeys*. A journey can be defined as follows:

Definition 3. *A journey J_i is a 4-tuple $(d_{i-1}, a_i, l_{i-1}, l_i)$ where d_{i-1} is the departure time at which the journey starts, a_i is the arrival time at which the journey ends, $l_{i-1} \in L$ is the origin location of the journey and $l_i \in L$ is the destination location of the journey.*

Suppose that we know the activity sequence of a certain individual. The sequence of journeys of the same individual will then have the following structure:

$$J_1 = (?, a_1, ?, l_1), J_2 = (d_1, a_2, l_1, l_2), \ldots, J_n = (d_{n-1}, a_n, l_{n-1}, l_n)$$

The chronological order defined on the a_i's and d_i's also applies when we consider a sequence of journeys. Given a sequence of activities, we do not know the precise departure time and origin of the first journey, which is indicated with the question mark symbol in the above sequence.

The relationship between journeys and activities can be exploited in this chapter. Although smart card datasets typically contain information on journeys, we can transform a sequence of journeys to a sequence of activities and vice versa. This can be done in a streaming fashion, as we only need to keep two journeys (or activities) in memory to produce an activity (or journey) assuming they are ordered in time.

6.3 Synthesis of Artificial Smart Card Data

One of the complications that rises in research related to mobility data in general and smart card data in particular, is the availability of the data. Real life data sets are often considered of significant importance by companies that collect and own the data, because of a competitive advantage or the sensitive nature of the data in relation to privacy. As a result many researchers have difficulties obtaining such data. In order to develop and test ways to deal with the data, it is important to have such data in the first place, but the best way to convince companies that they should provide access to the data is to show them that the developed methodologies will be indeed beneficial to them and their customers.

A second issue is that too much reliance on these company-owned data sets is detrimental for the scientific process, as it makes it extremely difficult to reproduce experiments. Furthermore, as real life data is not based on any model, it makes validation of proposed methodologies difficult due to a lack of a ground truth. In this section we provide an alternative to company-owned data sets by the introduction of a framework that can generate synthetic data sets that can be shared and discussed by researchers.

The framework uses *activity type* definitions and *generator* definitions to generate sequences of journeys and activities for individuals over any desired number of days. It assumes that an individual has a *home* activity type that serves as a starting point and ending point for a tour of activities connected by journeys. The performed activity types are selected using a Markov chain, while the choice of time, duration and location of each individual activity is dependent on the activity types.

A generated tour always starts at the home activity. For the current day, it is then decided if there are potential starting activities to be performed or whether the individual decides to stay at home for the day and try again the next day. If a starting activity is selected, the location and time for the first activity is then determined. For consecutive activities during the same tour, the starting time of the next activity is based on the ending time of the previous activity plus travel time. If at some point the home activity is selected as the next activity, the tour ends. This procedure also implements a number of rules that we will discuss in further detail in the appropriate sections.

6.3.1 Activity Types

In order to generate chains of activities, we define generators that are able to produce tours of activity types. We then fill in the location for each of these activity types on the generated tour. By design activity types play a central role in how the data that will be generated. An activity type declaration contains information related to the temporal properties of related activities, as well as possible locations at which the activity can be performed. For example, we may define one activity type for being at home, with possible locations in a residential area and within the city center, and define a second activity type work that can occur in the city center or in an industrial area. For every activity type, our definition needs to specify the following properties:

- A unique identifier. Every activity can be identified by a unique name.
- A starting time distribution. If the activity is selected as the first activity of a tour, the starting time is sampled using the specified distribution. If an activity never occurs as the first non-home activity in a tour, this distribution is never used in the sampling process.
- A duration distribution. In order to determine how long a non-home activity is performed, durations are sampled from this distribution.
- The days of the week during which this activity can begin. Some activities may only be available at certain days of the week. Work and school activities are typically only available from Monday until Friday, while activities such as events and music festivals are typically only available in the weekends.
- Whether an individual chooses a fixed location to perform this activity type. If the location has the *fixed* attribute, the same individual always performs this activity at the same location (although the location may be different for different individuals). Typical activity types with a fixed attribute are home, work and school, while shopping or entertainment are examples of activity types without fixed locations.
- The skip probability of this activity type. This is used to determine whether an individual leaves home in order to perform activities. Suppose the only activity type that is considered by an individual for a certain day has a skip probability of 0.3. Then with probability 0.3 that individual stays at home that day. If multiple activity types are considered at the beginning of the day, we define the lowest skip probability among those activity types to determine the probability to stay at home or not, but other implementations are possible.
- A set of possible locations L at which this activity type can be performed.

When an activity is generated for a certain individual, it is also necessary to determine the location for that individual. For this purpose, a set of possible locations is associated with each activity type. If an activity lacks the fixed attribute, a location is selected randomly each time any individual wants to perform this location. If it has the attribute, a new location is selected randomly only when a new individual performs the activity for the first time.

For each location l in the set L of an activity type definition, we define a number of attributes. Note that each activity type has its own set of locations, and that the same location can be defined for multiple activity types.

- A name useful for identifying the location.

- A latitude/longitude pair p_l to indicate the exact position of this location on the globe for the purpose of computing travel times.
- A weight w_l, useful to indicate that some locations are more popular for this activity type than others.

In order to determine the location for an activity type, we have two procedures. The first procedure is the most simple, as it only depends on the weights of the locations. The probability $p(l)$ to select location l is defined to be proportional to the weight of the location related to the other locations. Formally this is defined as follows:

$$p(l) = w_l \left(\sum_{k \in L} w_k \right)^{-1} \tag{6.1}$$

In certain cases the individual for whom we are selecting a location for is already at a different location o, which may influence the probability to select another location. For this, we use a distance function $d(o, l)$ which estimates how much time is required to travel from o to location l (our basic estimate takes 50 km/h as the crow flies, but more elaborate functions are possible). This distance function is used together with a logit model with a parameter λ. The probability to select a location l is defined as

$$p(l) = e^{-\lambda \frac{1}{w_l} d(o,l)} \left(\sum_{k \in L} e^{-\lambda \frac{1}{w_k} d(o,k)} \right)^{-1} \tag{6.2}$$

For practical purposes, we always use Equation 6.1 to select a location in case $\lambda \leqslant 0$. In case $\lambda > 0$, we use Equation 6.2 in case a previous location is available. If no previous location is available, such as when determining a home location, Equation 6.1 is used.

6.3.2 Generators

When all the activity types are declared, the second part of the model defines what types of activity chains will and will not occur. For the purpose of determining the sequence of activities that are performed by an individual, we use Markov Chains in a similar way as they are used to generate semi-realistic random natural language sentences. Instead of using only a single Markov chain to generate chains of activities, our procedure derives multiple Markov chains from the input data. As the input data contains blue prints instead of exact Markov chains, we use the more intuitive term *generators*. We define the following properties for a generator:

- A weight, indicating how likely it is that a new individual uses this generator to create a chain of activities and journeys.

- A home activity type, that is used as a starting and ending activity for every tour generated by this generator. In many applications this should be an activity type with a fixed location, although this is not strictly necessary.

- A list of transitions. It is highly recommended that these transitions are defined such that the home activity can be reached from any other activity and that all cycles in the Markov chain contain the home activity. This is to ensure that any started tour returns to the home activity after a finite number of steps.

Every time we start creating a chain of activities for a new individual, we first decide at random which of the generators to use with probabilities proportional to their defined weights. The Markov chain associated with that generator is defined by the transitions and the properties of the activity types. All transitions that are not defined explicitly are assumed to have a probability of zero. The transitions that are defined have the following properties:

- The *from* activity type, indicating from which current activity this transition is to be considered during the generation of journeys.

- The *to* activity type, indicating which activity is performed next when this transition is selected.

- The weight of the transition. The probability that a certain transition is selected is proportional to the weight of a transition among the relevant transitions that share the same *from* activity type.

Note that although a single Markov chain is defined, not all activities may be available every day. As a result, the actual transition probabilities to go from one activity type to another depend on the availability of that activity during the day for which we are currently generating activities.

Suppose that we have a set of activity types A and transition weights t_{ij} for $i \in A, j \in A$. The probability that activity j is selected to succeed activity i during the current day d is defined as follows:

$$p_d(i,j) = \begin{cases} t_{ij} \left(\sum_{k \in A_d} t_{ik} \right)^{-1} & \text{if } j \text{ available during day } d \\ 0 & \text{otherwise} \end{cases} \tag{6.3}$$

Here, A_d is the set of activity types that are available during day d. Note that during some days, it could be the case that there are no valid transitions. In those cases, the current activity is extended until a day is reached where a valid transition exists.

6.3.3 Sampling Process

Using all the properties defined in the previous sections, we can generate random data. For a certain individual, we generate T tours, starting from a given point in time, advancing the days until the individual returns to the home activity T times. The following procedure can then be repeated as often as desired to obtained the desired amount of data:

1. Pick a generator to use with probability proportional to the weights defined for the generators.
2. Repeat the following until the specified number of tours T has been generated
   a) Determine a day during which the new tour starts. First try the day after the day the last home activity started. If there are no eligible activity types for this day or all eligible activities for the new day have a skip probability greater than zero and a random draw is lower than the minimal skip probability, continue to the next day.
   b) Determine the next activity and determine the location to perform it. If the activity type has the fixed location attribute, reuse the earlier choice if available, otherwise pick a random location with probabilities proportional to their weights.
   c) Sample the starting time for the next activity and end the current home activity just in time to travel to the next activity in time.
   d) While the activity type of the current activity is not equal to home, repeat:
      i. Draw a duration for the current activity type and perform it for that duration.
      ii. Use the generator to determine the next activity type and its location.
      iii. Travel to the next activity type directly after the end of the current activity.
      iv. Start the next activity type as soon as you arrive at the new location.

6.4 Labelling of Activities

In this section we propose how to group stations based on observed time patterns while processing a database of journeys, convert the journeys to activities, and assign labels to the activities based on their temporal attributes. Using these labels, stations can be clustered to provide us with information how similar stations are. If we know the activity types associated with our dataset, we can use these clusters to investigate to what extent temporal properties can help us to identify what passengers are doing at a certain location and discover which stations are similar based on the activity types performed at their location.

6.4.1 Classification

One of the main goals of our methodology is to group locations that are similar based on the temporal distribution of activities. For this purpose, we introduce two types of *classifiers*, functions which map the space of some property of the activities to a discrete set of labels. We assume an analyst can manually define the *temporal* classifier, and the output of our clustering method can be used as a k-nearest neighbour classifier to obtain a *temporal* classifier. The formal properties of these classifiers can be defined as follows:

- The *temporal* classifier, which takes the starting and ending time of an activity and produces a label based on the temporal properties of the observation. We call the set of possible labels C^T.
- The *spatial* classifier, which takes the location where the activity takes place and produces a label based on its location. We call the set of possible labels C^L. The output of a clustering algorithm can be used to classify activities based on their location.

Starting with a sequence of journeys we can transform these journeys into a sequence of activities and use one of the classifiers to label the individual activities. This way, we obtain a *labeled activity sequence*. As long as the sequences of activities are already sorted, we can do this transformation in a streaming fashion, i.e. we do not need to store the entire smart card data set in the random acccess memory of a single computer, but are able to process the activities as they come.

Defining a temporal classifier is relatively straightforward: We decide a level of granularity and the temporal patterns that are relevant to us and construct a decision tree that corresponds to these choices. We propose a *time classifier* that only looks at the starting time of an activity as well as a *duration classifier* that considers the duration of the activities. The exact classification rules we use in this chapter are based on the work of Chapter 5 and are displayed in Figure 6.1. We can combine the labels of these two classifiers to obtain labels with both time and duration, and also expand the classifier to include whether the activity takes place during the week (Monday to Friday) or in the weekend (Saturday or Sunday).

Although many combinations are possible, we will consider results for three temporal classifiers: The combination of the start time classifier and the duration classifier (producing labels such as "Short Afternoon"), the combination of the start time classifier and the week/weekend classifier (producing labels such as "Week Noon") and the combination of the duration classifier and the week/weekend classifier (producing labels such as "Weekend Long"). Note that there is an important trade off to make here: Have a greater details in the classifier provides more information to the subsequent steps in the processing pipeline, but also make analysis of the results more complicated as the generated distributions will contain more data.

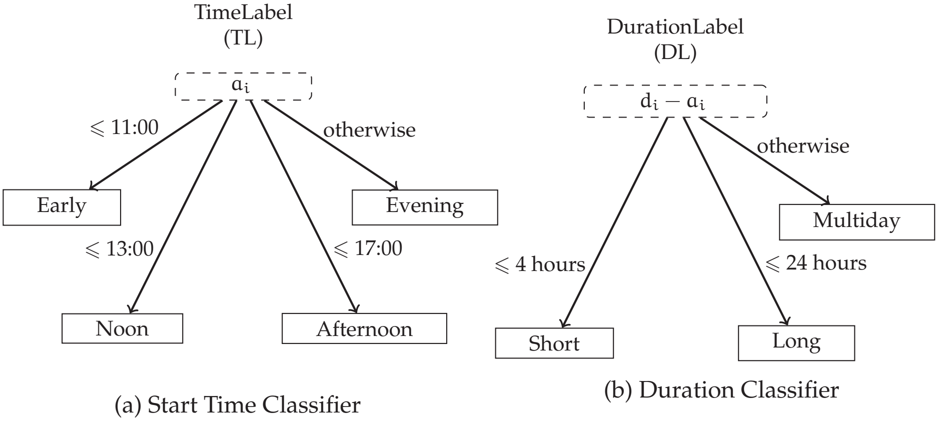


Figure 6.1: Two possible temporal classifiers. They can be combined or expanded to create derivative classifiers.

Table 6.2: Conversion from label occurrences (counts) to probabilities per location

(a) Count Matrix

Temporal Label	l_1	l_2	...
Short	5	6	...
Long	2	12	...
Multi day	3	2	...

(b) Probability Matrix

Temporal Label	l_1	l_2	...
Short	0.5	0.3	...
Long	0.2	0.6	...
Multi day	0.3	0.1	...

Defining a *spatial* classifier is a bit more complicated. While it is possible to create groups of locations and classify locations manually, a different approach is to cluster locations based on their *temporal* properties. First, we consider all the activities that have been observed at the individual locations and label them using the temporal classifier. We create a matrix of the observed frequencies of the labels, which will have a structure similar to Table 6.3a. Then, we normalize this matrix over the column to create a matrix that contains the probability that a certain temporal label is observed at that station as is given in Table 6.3b, based on the matrix with the counts.

The matrix with normalized label probabilities (as displayed in Table 6.3b) is suitable for a clustering approach as discussed in Appendix 4. We use k-means clustering in our experiments with Euclidean distances between the discrete distribution vectors of the locations, but other clustering approaches can be considered as well. In order to pick the number of clusters k that is required k-means clustering, we compute a scree-plot (which is explained in Appendix A). This plots the relation between the sum of the *within cluster distances* against the number of clusters k. Preferably we choose a value k after which the within cluster distance decreases slowly compared to earlier values. In case of doubt, a lower value of k is preferred as it makes interpretation of results simpler.

The results of k-means clustering can also be used to obtain a *spatial* classifier. The core idea of k-means clustering is to select a centroid (or mean) of each cluster, and assign every point to the centroid which it is closest to with respect to the Euclidean distance.

6.4.2 Validation

In order to assess the quality of the clustering method we can compute to which extent each cluster of locations matches the locations associated with each activity type. Let us consider L to be the set of all locations and $L_1, L_2, \dots, L_k$ the output of our clustering method. Let us now consider the A the set of defined activity types used to generate the data set, and for each activity type $a \in A$ we have L_a as the set of locations at which activity type a can be performed. We can compute for each pair of an activity type a and a cluster $i \in \{1, \dots, k\}$ the overlap of L_a and L_i. We define the overlap fraction o_{ai} of such a pair as follows:

$$o_{ai} = \frac{|L_a \cap L_i|}{|L_a|} \tag{6.4}$$

In an ideal situation we would have a perfect match between an activity type a and a cluster i. If we create a table, this situation would result in at most o_{ai} values to be 0, while some are 1. To asses the quality of the clustering output and validate whether it represents the original patterns well, we can inspect a table of overlap fractions. Note that the ideal situation is not likely to occur in practice, as it requires the number of clusters to be equal to the number of activity types. For actual data it is not likely that we can define the exact number of activity types. Furthermore, the ideal situation requires that there is a one-on-one correspondence between activity types and locations. As at least some stations are likely to offer multiple activity types, it is expected that at least some overlap fractions will be between 0 and 1. In those cases, it is better if the number of non-zero entries in each row and each column is small.

6.5 Synthetic Smart Card Data

In order to evaluate the applicability of our proposed methodology, we apply it to synthetic smart card data of which we known the underlying demand patterns for the purpose of validation. There are a number of reasons to work with synthetic data first. Synthetic data can be shared with other researchers and thus increase the transparency and reproducibility of our results. Furthermore, for real world smart

Table 6.4: The temporal activity types defined in the synthetic smart card data. This includes the distributions for the start time of an activity and its duration (n denotes a normal distribution, otherwise it is uniform), on which days the activity will be performed and whether the same individual will always perform the activity at a fixed location.

Activity	Start	Duration	Days	Locations	Fixed
home	-	-	All	25	Yes
work	n(9:00,30m)	n(8h,1h)	Monday – Friday	14	Yes
work-always	(8:00 – 23:00)	n(9h,1h)	All	2	Yes
work-evening	n(15:00,30m)	n(8h,1h)	Monday – Saturday	5	Yes
shop-day	n(13:00,2h)	n(30m,15m)	Monday – Saturday	5	No
shop-evening	n(17:00,1h)	n(30m,15m)	Monday – Friday	4	No
shop-major	n(11:00,1h)	n(5h, 90m)	Monday, Friday, Saturday	3	No
study	n(11:00,1h)	(2h – 6h)	Monday – Friday	4	No

card data, we have no objective way to compare the results of our methodology to a "ground truth" and as a consequence is not suitable for validation of a methodology.

For the generation of smart card data, we use the methodology introduced in Section 6.3. We developed a synthetic demand template for the city of Utrecht, in the Netherlands. The activities used in this demand template are introduced in Table 6.4.

We define different types of work activities and different types of shop activities. For example, work-always activities can start at any point during the day and thus models shifts in hospital or security, while the regular work activity is more like an office job. Similar distinctions are made for the shop activities, where quick shopping can be done during the day or in the evening, but major shopping can take a lot more time. More activity types can be added, but as we intend to validate our methodology we need to compare the results of our method against the different activity types. We prefer to keep the number of activity types limited to make sure this comparison is manageable.

In the demand template, we also specify potential locations for all activity types. We have selected a number of locations in the city that can typically be associated with shopping areas, important public transport hubs, university locations, hospitals and residential areas. We assume that people live at all locations, even if there are for example shops at that location, but that the purely residential areas without such activities are twice as likely to be the home of a person than the areas where other activities can take place. Besides, there is a large overlap between the locations where shopping activities can be performed and where people can work, as shops have to be operated by employees and are often near popular office locations. The number of shopping locations that are open in the evening is smaller than the ones that are open during the day. Because of this the list of shop-evening locations is smaller than the list of shop-day locations. Finally, the shop-major locations include some irregular

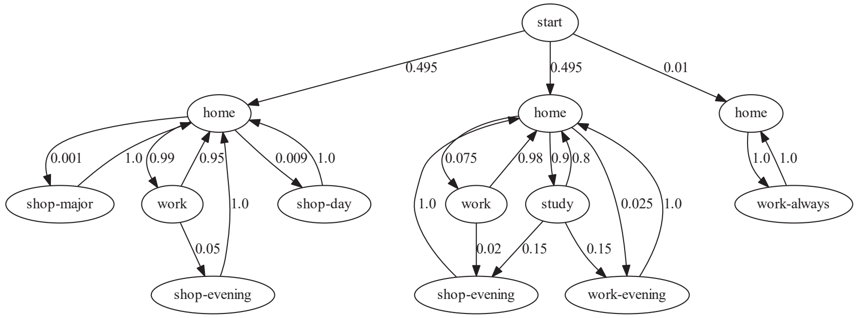


Figure 6.2: Graphical representation of the Markov process that is used to determine in which order different activity types are performed. The left part of the graph represents individuals who work and do different kinds of shopping. The middle part of the graph represents students who study and have their jobs often in the evening. The right part of the graph represents individuals who have to work on a wide range times and do nothing else.

locations. An area with a number of furniture shops and a large gardening shop are included under shop-major, but not under shop-day or shop-evening.

As a second step, we specify one or more Markov chains of the possible activity sequences. The structure defined for the synthetic demand template is shown in Figure 6.2. We define three behavioral archetypes: A working person who goes shopping once in a while, a student who also does shopping and prefers a job in the evening, and a person who only works at irregular times, such as nurses in a hospital.

We have used this model to randomly generate 200,000 smart card logs consisting of 50 tours that start and end at the home location of that smart card. This process takes roughly a minute on a computer with an i7 4770 CPU and results in 1.8 GB of textual data. The dataset is published on Zenodo (Bouman, 2017).

6.6 Results

As a first step, our software determines the number of clusters we use to classify the locations. The scree-plot utilized for this process is shown in Figure 6.3, which displays the results for the three temporal classifiers introduced in Section 6.4. From the scree plot we can decide that the case where we only consider the combination of arrival time and duration will be used to produce *seven* clusters, the classifier that combines duration with week/weekend will be used to produce *five* clusters and the

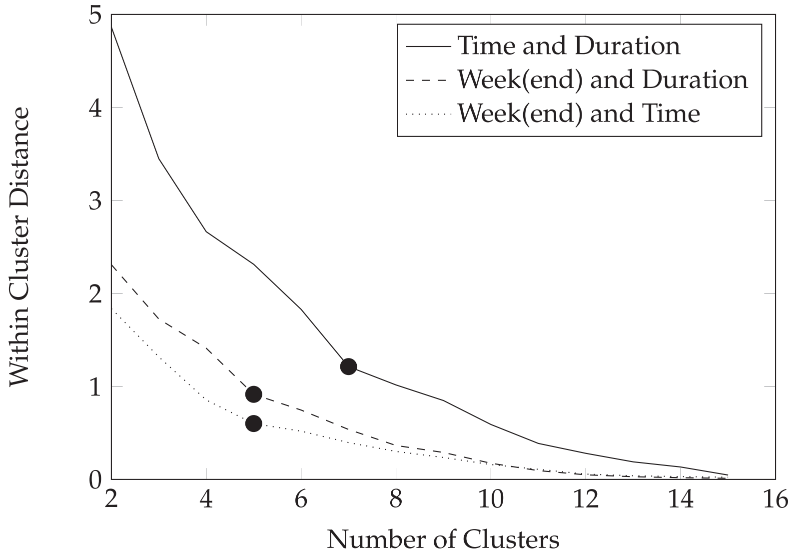


Figure 6.3: Scree plot for three different time classifiers applied to the synthetic smart card data. The number of clusters which was finally selected is indicated with a black dot.

classifier that combines arrival time with week/weekend will be used to produce *five* clusters. As the scree plot is relatively smooth, these choices are somewhat debatable, especially for the classifiers that include the week/weekend label. This indicates that there is room for future research related to the classification and clustering steps.

With the choices for the number of clusters per classifier, we can apply both steps to summarize the full data set using the computed cluster centroids. These are displayed in Table 6.5. We can interpret the clusters based on the frequency of the different labels in each centroid. For example, a cluster that has a high value for the “Overnight” label is likely associated with residential areas, while a cluster that has a high value for “Long Early” labels is likely associated with areas where people work. If both labels occur frequently within a cluster, it is an area where both people live and work.

In Table 6.6a we can observe a number of different cluster types. Clusters A3 and A5 contains mostly “Long Early” and “Overnight” activities, indicating that this cluster contain areas where people both live and work. Cluster A1 focuses purely on the “Overnight” activities, indicating that this cluster is associated with areas that are purely residential. Clusters A6 and A7 contain both “Long Early” and “Overnight” activities, but also a couple of short activities indicating that these are areas where people live, work and do some short shopping. Cluster A4 again contains “Long Early” and ”Overnight”, but also “Long Afternoon” and “Short Afternoon”, indicating that

Table 6.5: Location clustering results for three type of temporal classifiers applied to the synthetic smart card data. Locations were clustered based on the discrete distribution vectors of the different temporal labels of the activities observed at that location. Each row indicates the average distribution of activity labels for activities observed at locations in the cluster, as well as the number of stations covered by the cluster. For readability, some abbreviations are introduced, such as A.noon for afternoon and Eve. for evening. All labels that occur with value greater than 0.1 in the centroid are bold.

(a) Clustering based on a temporal classifier which produced labels for both duration and arrival time.

	Long					Short				
Cluster	A.noon	Early	Eve.	Noon	Overnight	A.noon	Early	Eve.	Noon	Stations
Cluster A1	**0,03**	**0,04**	0,00	**0,01**	**0,92**	0,00	0,00	0,00	0,00	11
Cluster A2	**0,01**	**0,13**	0,00	**0,12**	0,00	**0,02**	**0,37**	0,00	**0,36**	2
Cluster A3	**0,02**	**0,66**	0,00	**0,01**	**0,30**	0,00	0,00	0,00	0,00	6
Cluster A4	**0,38**	**0,23**	0,00	**0,02**	**0,17**	**0,13**	**0,01**	**0,03**	**0,04**	1
Cluster A5	**0,06**	**0,35**	0,00	**0,01**	**0,47**	**0,07**	**0,01**	**0,02**	**0,02**	2
Cluster A6	**0,03**	**0,49**	0,00	**0,01**	**0,15**	**0,18**	**0,03**	**0,03**	**0,07**	3
Cluster A7	**0,08**	**0,26**	0,00	**0,07**	**0,22**	**0,01**	**0,19**	0,00	**0,18**	2

(b) Clustering based on a temporal classifier which produced labels for week or weekend and the duration of activities.

		Weekend		Week		
Cluster	Overnight	Long	Short	Long	Short	Stations
Cluster B1	**0,22**	0,00	0,00	**0,77**	0,00	11
Cluster B2	**1,00**	0,00	0,00	0,00	0,00	2
Cluster B3	**0,62**	0,00	**0,01**	**0,35**	**0,02**	4
Cluster B4	**0,70**	0,00	**0,08**	**0,13**	**0,10**	2
Cluster B5	**0,79**	0,00	**0,02**	**0,18**	**0,01**	8

(c) Clustering based on a temporal classifier which produced labels for week or weekend and the arrival time at an activity location.

		Week				Weekend				
Cluster	Overnight	A.noon	Early	Eve.	Noon	A.noon	Early	Eve.	Noon	Stations
Cluster C1	**0,22**	**0,44**	0,00	**0,31**	**0,02**	0,00	0,00	0,00	0,00	11
Cluster C2	**1,00**	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2
Cluster C3	**0,70**	**0,15**	0,00	**0,10**	**0,01**	**0,02**	0,00	0,00	**0,01**	5
Cluster C4	**0,56**	**0,24**	0,00	**0,16**	**0,01**	**0,01**	0,00	0,00	**0,01**	2
Cluster C5	**0,79**	**0,11**	0,00	**0,07**	**0,01**	**0,01**	0,00	0,00	**0,01**	7

these areas can be used both for long and short shopping in addition to living and work activities. Finally, cluster A2 contains a mix of “Long Early”, “Long Noon”, “Short Early” and “Short Noon” activities, indicating that no people live in these areas. Note that although some clusters are associated with the same “types” of activities, the distribution of how frequent these activities occur can be quite different.

The second clustering output displayed in Table 6.6b is based on only the duration of activities and the week or weekend label. Only cluster B4 stand out because it only contains “Overnight” activities, while all other clusters focus on both “Overnight” and “Week Long” activities, but with different distributions. Cluster B4 contains most of the short activities. When one is only interested in which activity types occur at certain stations, this clustering output is not very informative.

The third clustering output displayed in Table 6.8a is based on the time at which activities take place and the week or weekend label. Cluster C2 captures stations that only have “Overnight” activities and thus correspond to residential areas. None of the centroids seem to have high values in any of the “Weekend” patterns, indicating that the activities with “Week” labels play a more significant role among the activity types. There are no early activities in the clustering outcome, while such activities were important in the clustering outcome of Table 6.6a.

For the purpose of validation, we consider the original model used to generate the data and check to what extent the clusters capture specific activity types defined in the smart card data generator. One station always has one or more activity types that can occur at that location, and we consider which percentage of the stations in a cluster coincide with the stations where a particular activity type is defined. Ideally, we have values of 100% and few strictly positive values per row, as this indicates that a cluster matches specific activity types very well. The results of this validation step are presented in Table 6.7

The classifier which combines arrival times and duration is able to distinguish certain activity types the best from the classifiers we tested. This can be observed in Table 6.8a: Cluster A1 captures only home activity locations, while cluster A2 captures only stations where study activities occur. Cluster A3 captures locations where shopping and work activities take place. Clusters A4 and A5 capture locations where many different types of activities take place and are not very informative as a result. Cluster A6 captures a large fraction of locations where shopping activities take place, while cluster A7 captures locations where study and irregular work activities are performed.

The clustering based on a combination of duration and week/weekend labels is lessactivity types reasonably well. In Table 6.8b we can see that cluster B1 captures locations where home activities take place, while cluster B2 captures locations where study activities take place. Cluster B4 captures locations where shopping activities take place reasonably well. Clusters B3 and B5 are harder to interpret and thus not as informative as the other clusters.

The clustering based on the classifier that combines arrival time and week/weekend shows a similar and can be seen in Table 6.8c. This suggests that the classifier that

Table 6.7: Validation based on the overlap fractions o_{ai} of combinations of the locations in each cluster i that correspond to the locations in the set L_a defined for a particular activity type a. If we see a percentage of 100%, this means that a certain cluster has captured all locations at which that activity type can be performed. If we see a percentage of 50%, half of the locations at which an activity can be performed is captured by the locations within that cluster. These percentages are given for three different types of temporal classifiers. The best situation is when a cluster captures the locations of as few activity types as possible, preferably with high values. Values with more than 40% are displayed in a bold font.

(a) Table based on the clustering based on a temporal classifier that generates labels based on duration and arrival times.

Cluster	home	shop-day	shop-eve.	shop-major	study	work	work-alw.	work-eve.
A1	**44%**	0%	0%	0%	0%	0%	0%	0%
A2	0%	0%	0%	0%	**50%**	0%	0%	0%
A3	24%	0%	0%	**67%**	0%	**43%**	**50%**	20%
A4	4%	20%	25%	0%	0%	7%	0%	20%
A5	8%	20%	25%	0%	0%	14%	0%	20%
A6	12%	**60%**	**50%**	33%	0%	21%	0%	20%
A7	8%	0%	0%	0%	**50%**	14%	**50%**	20%

(b) Table based on the clustering based on a temporal classifier that generates labels based on week or weekend and the duration of activities.

Cluster	home	shop-day	shop-eve.	shop-major	study	work	work-alw.	work-eve.
B1	**44%**	0%	0%	0%	0%	0%	0%	0%
B2	0%	0%	0%	0%	**50%**	0%	0%	0%
B3	16%	20%	25%	33%	0%	29%	0%	20%
B4	8%	**40%**	**50%**	0%	0%	14%	0%	20%
B5	32%	**40%**	25%	**67%**	**50%**	**57%**	**100%**	**60%**

(c) Table based on the clustering based on a temporal classifier that generates labels based on week or weekend and the arrival time at activities.

Cluster	home	shop-day	shop-eve.	shop-major	study	work	work-alw.	work-eve.
C1	**44%**	0%	0%	0%	0%	0%	0%	0%
C2	0%	0%	0%	0%	**50%**	0%	0%	0%
C3	20%	**40%**	**50%**	33%	0%	36%	**50%**	**40%**
C4	8%	20%	25%	0%	0%	14%	0%	20%
C5	28%	**40%**	25%	**67%**	**50%**	**50%**	**50%**	**40%**

combines duration and activity time yields more informative results than a classifier that combines a single one of these options with a week/weekend label.

6.7 Discussion and Future Research

We have shown that we can generate a large amount of synthetic smart card data. We can analyze the temporal are spatial properties of activities implied by the smart card data and make groups of similar stations, all within reasonable time. As the synthetic smart card data is generated from a predefined set of demand patterns, we can validate the quality of the obtained clusters of stations. We learn that certain clusters match specific activity types by the tested approach, while other activity types are more difficult to distinguish. This makes sense, as some activity types may be similar, such as studying and working activities, while other locations offer different activity types at the same time, such as shopping, working and home activities.

From these results, we can reason that public transport operators do not know "everything" based on the smart card data. Even if there is a good way to segment passengers or activities, it is likely that some different activity types share temporal and spatial attributes and therefore will always be confused by looking solely at the smart card data. This relates strongly to the observation made in Chapter 4 that smart card data do not contain all information required to regard it as activity based demand.

The results show that although in theory smart card data miss information, it undeniably contains more information than traditional ticket sales data or randomized manual passenger counts. Since some clusters could be associated with specific activity types, such as shopping or studying, there is fertile ground to develop methodologies that allows operators to explore the goals passengers pursue with their travels. The approach of generating synthetic smart card data and validating the outcome of such methods proposed in this chapter will be a valuable tool in the development of future methods.

There are a number of ideas that can be considered for future improvement of the methodology tested, or even for alternative methodologies. Since we make use of basic k-means clustering it seems there is room to experiment with other data mining techniques. One idea could be to work directly with the temporal attributes of the activities instead of the labels we currently work with, or with a hybrid approach that assigns a likelihood to each label instead of picking a single label. This keeps more information available during the clustering process, which could potentially be exploited by more sophisticated clustering methods, but this makes it more challenging to present the output concisely to the end user.

Another idea to consider for future research is to use thresholds in the processing pipeline to produce binary vectors, indicating whether a certain type of label occurs with a high enough frequency at a station, instead of the distribution vectors we currently used. This would result in greater similarity for a station that has both residential and working activities, even if the frequency of these activities is different. This could result in the need for fewer clusters, but at the cost of information related to the distribution of the activity types in the final summary.

Part III

Individual Decision Strategies Under Uncertainty

7

Passenger Route Choice in Case of Disruptions

Co-authors : Marie Schmidt, Leo Kroon, Anita Schöbel
This paper has been accepted after peer review for presentation at and publication in the proceedings of IEEE Conference on Intelligent Transport Systems, IEEE-ITSC2013.

7.1 Introduction

One of the major nuisances a passenger in railway transport can experience is a disruption, resulting in the cancellation of a number of train services. Not only do disruptions usually cause significant delays of the passengers' journeys, but they may also confront the passenger with a complicated question: "*What is the best way to continue my journey?*" In such a situation, a little bit of information may be extremely valuable for making the right choice. As such, it is important that operators do the best they can in informing their passengers. However, even the operator himself may not have all information that would be required to give the best possible advice. In many cases it can take some time to assess the cause and severity of a disruption. In other situations the cause may have been resolved but the exact time before operations are back to normal is still uncertain.

In case good information is not available, the passenger faces a dilemma. Should he be optimistic about the duration of the disruption and hope it is over soon enough to wait for the planned fast connection, or should he be pessimistic and take an alternative that might take much longer than the disrupted connection? Both approaches may lead to an unfortunate outcome, as the optimistic passenger might wait much longer than anticipated before the disruption is over, while the pessimistic passenger may find out that the disruption had vanished just after he departed on his lengthy detour.

In the past decades, algorithms for situations where the input arrives over time, have been investigated under the name of "online algorithms". Often the performance is measured by the so-called "competitive ratio", which can be interpreted as the worst-case ratio of the obtained solution value to the best achievable solution value when having the required information in advance.

In this chapter we study the quality of strategies in an online passenger waiting problem, where a passenger has to decide between waiting for the end of the disruption or taking a detour. We characterize the strategies of the passenger and analyze the competitive ratios of these strategies. We apply the results to some realistic disruption scenarios in order to show the applicability of this approach in decision support systems that can improve passenger guidance in a disrupted situation. Additionally, we investigate a version of the problem where we perform an average case analysis using different probability distributions.

7.2 Related Work

Over the years, many approaches dealing with decision making under uncertainty have been developed.

One popular way of analyzing online algorithms is *competitive analysis*, where the worst case ratio of the online solution compared to the offline solution is used to measure the quality of an online strategy (Sleator and Tarjan, 1985; Borodin and El-Yaniv, 1998). Another way of analyzing solution strategies is to introduce a probability distribution on the uncertain input parameters and consider the expected value of the solution value. An example of this approach is Fujiwara and Iwama (2002). In this chapter we refer to this approach as *average-case analysis*.

The first competitive analysis of online strategies in the literature appeared in Sleator and Tarjan (1985), where the performance of different online update rules on a list data structure was compared to the optimal offline schedule, i.e., the best schedule if all information was known before.

Other frameworks to deal with uncertainty are robust optimization (Ben-Tal et al., 2009), which aims to find a solution which is feasible for all possible realizations of scenarios, and stochastic optimization (Birge and Louveaux, 1997), where a probability distribution on scenarios is given and a solution needs to be feasible with high expectation.

Several uncertain shortest path problems on networks have been considered in the literature. For example, the *Canadian traveler problem* (Papadimitriou and Yannakakis, 1991; Bar-Noy and Schieber, 1991) asks how to find a path through a network in which edges may be blocked, which is revealed to the traveler when an adjacent node is reached. A different class of path finding problems deals with the situation where edge-lengths are uncertain (Yu and Yang, 1998). This problem has been studied under the aspect of regret-robustness, i.e., minimizing the worst-case difference between the length of the chosen path and an optimal path over all scenarios. In Büsing (2009), travel advice is given to passengers by providing them with a sub-network that contains a shortest path for each scenario from two uncertainty sets (the interval-based and Γ-uncertainty set).

However, the approaches used there are not directly applicable to uncertain routing problems in railway transport. One approach to give travel advice in case of uncertainty is made in Goerigk et al. (2013a), where robustness of passenger routes against train delays is studied, mainly focusing on passenger transfers.

In Tsuchiya et al. (2006) a decision support system for passengers in a disrupted public transport system is discussed, where the travel time of different routes is predicted using a "resumption process model" fitted on past data of disruptions. It also deals with how the advice can be presented to the user.

In this chapter, we also consider route choice in railway transport systems under uncertainty. We assume that there is a disruption of uncertain length on a specified route. We want to give travel advice to passengers who planned to travel on this route. An important aspect of our approach is that possible routes depend on an underlying periodic timetable. We evaluate different travel strategies with respect to the competitive ratio and the expected arrival time.

7.3 Passenger Route Choice in Case of Disruptions

Let us consider the case where a passenger wants to make a journey using a certain railway transport service. He arrives at his departure station at time 0 and discovers that his preferred (and fastest) connection is obstructed by a disruption. No information about the cause or duration of the disruption is given, but an alternative connection is available. Since it is usual in railway transport that different connections run multiple times during a day, the passenger can actually decide not to take the first departure available, but wait for a later connection and see what happens to the disrupted connection meanwhile. Although we assume the goal of the passenger is to arrive at his destination as early as possible, under certain circumstances it can be beneficial to wait. For example, if the alternative connection departs in 8 minutes and the next regular connection departs in 12 minutes, it could be the case that the disruption is resolved in 10 minutes. If the alternative connection is at least 5 minutes slower than the regular one, the passenger would benefit from waiting in this situation. However, the problem is that the passenger does not know that the disruption will be resolved before the next regular connection departs at the moment he has to decide whether to board the alternative connection.

In order to analyze this problem, we first introduce some notation by means of a formal problem definition:

Definition 4 (Passenger Waiting Problem)**.** *An instance of the* passenger waiting problem *consists of a service schedule* $(\Delta S, S, T_S, y)$ *for a disrupted regular connection and a service schedule* $(\Delta D, D, T_D)$ *for an alternative connection. In these schedules,* ΔS *and* ΔD *denote the first scheduled departure time (relative to the arrival of the passenger at time* 0*) for both services respectively, and* S *and* D *denote the travel duration of both connections. The two services are repeated regularly after periods of* T_S *and* T_D *units of time. Furthermore,* y *denotes the time the regular connection needs to travel until it hits the disruption (therefore* y *is restricted to* $0 \leqslant y \leqslant S$*). Finally, we denote the time at which the disruption is resolved by* x*.*

To calculate the departure times of the next trains departing after any given moment t *we define the functions*

$$\mathrm{suc}_S(t) = \Delta S + \max\left(0, \left\lceil \frac{t - \Delta_S}{T_S} \right\rceil\right) T_S$$

which gives the scheduled departure time of the next train on the regular connection after the moment t*, and*

$$\mathrm{suc}_D(t) = \Delta D + \max\left(0, \left\lceil \frac{t - \Delta_D}{T_D} \right\rceil\right) T_D.$$

which is the departure time of the next train on the alternative connection after the moment t*.*

When we know the decision of the passenger, we can calculate his arrival time based on the time the disruption is resolved, x*. If the passenger decides to board the* i*th departure of the alternative connection, he will thus arrive at time* $\Delta D + iT_D + D$*. If the passenger boards the* i*th departure of the regular connection, he will arrive at the time the first regular connection that goes unhindered through the disruption will arrive. This is at time* $\max\{\Delta S + iT_S + S, \mathrm{suc}_S(x - y) + S\}$*.*

Assuming reasonable passenger behavior, we can assume $\Delta D + D > \Delta S + S$. We also assume there is no ambiguity in the order in which departure events happen. If events would coincide, we should adapt the schedules by adding an ϵ to either ΔS or ΔD. Although it is possible to work without this assumption, it makes the formulas less complicated and does not really influence our arguments and results.

The Online Passenger Waiting Problem

Let us consider an instance I containing both the regular schedule $(\Delta S, S, T_S, y)$ and the schedule of the alternative connection $(\Delta D, D, T_D)$. For the online setting we assume that a passenger does not know the value of x, but receives a message immediately when the disruption is over and never earlier. As a result the passenger will not know x before it is actually time x.

Example Instances

Let us consider the map in Figure 7.1 showing part of the passenger railway services operated by Netherlands Railways (NS). Suppose a passenger wants to travel from Gouda to Rotterdam. There is a disruption on the track between Gouda to Rotterdam and therefore no trains bound for Rotterdam are leaving Gouda. However, the tracks to The Hague (on the map: Den Haag) are still accessible and thus the passenger has the option to make a detour. The regular connection usually departs every 15 minutes and has a duration of 10 minutes. At the moment the passenger arrives at Gouda, the next scheduled departure is within 5 minutes. This gives us $\Delta S = 5, S = 10, T_S = 15$. We assume the first train to The Hague departs in 1 minute. The detour through The Hague takes roughly 50 minutes in total and departs every 15 minutes from Gouda. This gives us $\Delta D = 1, D = 50, T_D = 15$. Since no direct trains for Rotterdam are leaving Gouda, we have $y = 0$.

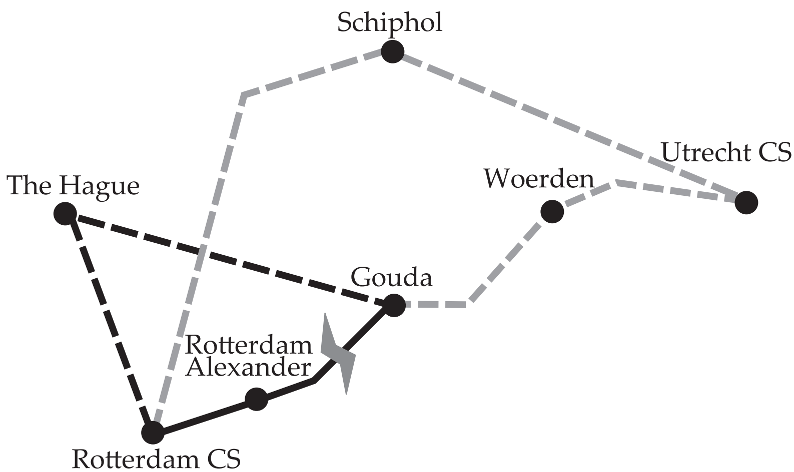


Figure 7.1: Example scenario for a passenger traveling from Gouda to Rotterdam

In our second example we consider a case where some distance (and hence some time) must be traveled before arriving at the disrupted connection, thus $y > 0$. Let us consider a passenger who leaves from Utrecht to travel to Rotterdam. The regular connection goes through Gouda as well and normally takes about 30 minutes. Now suppose the disruption between Gouda and Woerden, so there are no trains bound for either Rotterdam or The Hague leaving from Gouda. In such a situation, Netherlands Railways (NS) usually advises passengers to travel via Schiphol (which takes about 85 minutes). But as Gouda is on the regular connection from Utrecht to Rotterdam and it is likely that there are still trains bound for Gouda that leave from Utrecht, the passenger can make a gamble and travel towards Gouda, hoping that by the time he reaches Gouda the disruption will be over. The map of this scenario is shown in Figure 7.2.

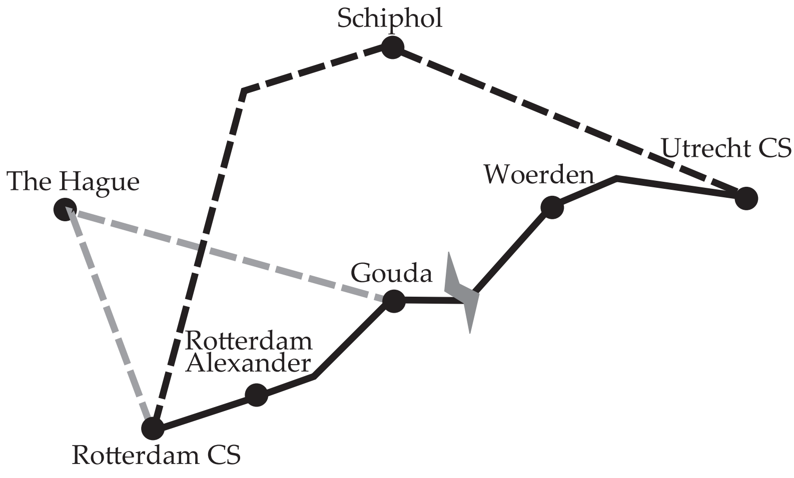


Figure 7.2: Example scenario for a passenger traveling from Utrecht to Rotterdam

In this case, the regular connection has $S = 30, T_S = 15$ and for example $\Delta S = 5$. However, as it takes 20 minutes to reach Gouda, we get $y = 20$. The detour through Schiphol usually takes $D = 85$ minutes and departs every $T_D = 30$ minutes from Utrecht, the first one for example at time $\Delta D = 2$. As we can see in Section 7.7, the scenarios with $y = 0$ behave differently from the scenarios with $y > 0$.

7.4 Competitive analysis of the online passenger waiting problem

The concept of competitive analysis as discussed in Section 7.2 was introduced by Sleator and Tarjan (1985). One possible approach to an online analysis which we will use in this chapter is the game-theoretic interpretation. We consider a one-round game where we have two players: An algorithm player who picks an algorithm $\mathcal{A}$ and an instance player or adversary who afterwards picks an instance I. The algorithm player wants to perform as well as possible, i.e., to *minimize* the ratio of $\mathcal{A}$ applied to I and an optimal solution of I in case of full information. The adversary, on the other hand, wants to *maximize* this value. If the adversary chooses the best-response to the algorithm $\mathcal{A}$ played by the first player, we obtain the worst-case competitive ratio of $\mathcal{A}$ as outcome of the game.

The Competitive Ratio

In general, if we consider a set of all possible instances $\mathcal{I}$, the competitive ratio of a certain algorithm $\mathcal{A}$ is given by

$$C^{\mathcal{A}} = \max_{I \in \mathcal{I}} \frac{\text{online}_{\mathcal{A}}(I)}{\text{offline}(I)}.$$

Here, offline(I) is the best possible solution value for instance I in an offline setting where all instance data (in our case also the time instant x the disruption is over) is known when the best decision is calculated, and online$_{\mathcal{A}}(I)$ denotes the solution value of the algorithm applied to instance I.

Now let us consider the competitive ratio of the online passenger waiting problem. We will analyze the decision strategies of the passenger using a competitive analysis. For a given decision strategy of the passenger, it is not difficult to construct an instance with arbitrarily large competitive ratio. As long as the adversary has the flexibility to pick a schedule, he can pick T_S and T_D very large and pick y and x such that the algorithm always will end up in the wrong connection.

Hence, we cannot provide passengers with an easy strategy which works well for all disruption situations. However, for a given instance I of the online passenger waiting problem, it is possible to determine the strategy with best competitive ratio as we see in the following. By I_x we denote an offline version of the problem where the time x at which the disruption is over is fixed.

If we have a set of strategies indexed by r, then the online performance of a strategy r on instance I for a given value of x is denoted by online$_r(I, x)$. For such a strategy r, the competitive ratio on instance I is defined by:

$$C_I^r = \max_{x \geqslant 0} \frac{\text{online}_r(I, x)}{\text{offline}(I_x)}.$$

Again, we can interpret this from a player-adversary point-of-view. Let the passenger pick a strategy r. In reply to this, the adversary may now pick an x in such a way that the performance of the strategy is as bad as it can be. The competitive ratio of strategy r on an instance I is attained by taking the maximum of the performance ratio over all possible values of x.

r-Greedy strategies

It is likely that real passengers do not decide whether to wait by flipping a coin and thus use deterministic strategies in these cases. Because of this, we will ignore randomized strategies. Since x is not bounded from above, any strategy that makes the passenger board the disrupted train before the disruption is over has an unbounded competitive ratio. Just the same, every strategy which does not specify a waiting time limit after which the passenger boards alternative train if the disruption has not vanished so far has unbounded competitive ratio. Since alternative trains depart only at time instants $\Delta_D + rT_D$ for the integers r, we can restrict ourselves to consider only strategies which do the following:

- take the train which arrives earliest at its destination, if the disruption is over before time $\Delta_D + rT_D$,
- take the alternative train then, if the disruption is not over at time $\Delta_D + rT_D$.

We will call such a strategy an r-greedy strategy.

Calculating the competitive ratio of an r-greedy strategy on a certain schedule instance

Now that we have introduced the notion of the competitive ratio and the notion of the r-greedy strategy, we will show how we can efficiently calculate the competitive ratio of such a strategy on an instance of the online passenger waiting problem where the schedules are fixed, but x is not.

Theorem 7.1. *The competitive ratio* C_I^r *of an* r*-greedy strategy on an instance* I *is the maximum of*

$$\frac{\Delta D + rT_D + D}{\min\{\mathrm{suc}_S(\Delta D + rT_D - y) + S, \Delta D + D\}} \tag{7.1}$$

and

$$\max_{x \in C_1 \cup C_2 \cup C_3} \frac{\min\{\mathrm{suc}_S(x) + S, \mathrm{suc}_D(x) + D\}}{\mathrm{suc}_S(x - y) + S} \tag{7.2}$$

with

$$\begin{aligned}
C_1 &:= \left\{\Delta_S + iT_S : i \leqslant \left\lfloor \frac{\Delta D + rT_D - \Delta S}{T_S} \right\rfloor\right\}, \\
C_2 &:= \left\{\Delta_S + iT_S + y : i \leqslant \left\lfloor \frac{\Delta D + rT_D - \Delta S - y}{T_S} \right\rfloor\right\}, \\
C_3 &:= \{\Delta_D + iT_D : i \leqslant r\}.
\end{aligned}$$

Proof. The fraction in (7.1) represents the maximum competitive ratio over the range where the disruption is over after the passenger has departed on a detour, so where we have $x > \Delta D + rT_D$. In that situation, $\mathrm{online}_r(I, x) = \Delta D + rT_D + D$ and therefore does not really depend on x anymore. The best decision for the adversary is thus to pick x in such a way that $\mathrm{offline}(I_x)$ is as small as possible and thus we take the limit of $x \to \Delta D + rT_D$, giving (7.1).

Note that we use the assumption that no departure events coincide and that we would need to add an epsilon to the suc_S expression to deal with such situations.

The case considered in (7.2) represents the maximum competitive ratio over the range where the passenger has not departed at the time the disruption is over. In this range we have $x \leqslant \Delta D + rT_D$, so we should only take the maximum over a limited interval. Within this interval, there are only a few number of possible values for x which influence the online or the offline strategies: The moments either the regular connection departs (C_1), the detour departs (C_3) or the regular connection reaches the disruption (C_2). We evaluate the online/offline ratio for each of these moments and take the maximum over all of them.

As (7.1) covers the case $x > \Delta D + rT_D$ and (7.2) covers the case $x \leqslant \Delta D + rT_D$, taking the maximum of both gives us the competitive ratio. □

Now that we have a way to calculate the competitive ratio of a given r-greedy strategy, we can calculate the value of r for which the competitive ratio is the smallest, which we call r^*. For this, it is sufficient to limit the range of candidate values for r^* using the following theorem.

Theorem 7.2. *If we have $\Delta D + rT_D > \Delta D + D$, then $C_I^{r'} \geqslant C_I^r$ for any $r' > r$.*

Proof. If the competitive ratio of C_I^r is attained using $x \leqslant \Delta D + rT_D$, then the same competitive ratio can be attained for $C_I^{r'}$. If the competitive ratio of C_I^r is attained using $x > \Delta D + rT_D$, the denominator for both r and r' will be $\Delta D + D$ and in that case we get $C_I^{r'} > C_I^r$. □

Now we can calculate r^* by evaluating

$$r^* = \operatorname{argmin}_{r=0}^{r_{\max}} C_I^r, \quad \text{where } r_{\max} = \left\lceil \frac{D}{T_D} \right\rceil.$$

7.5 Average case analysis of the online passenger waiting problem

Instead of taking the competitive ratio as a performance measure for our online strategies, in this section we compare strategies by an average case analysis, i.e., by comparing the expected arrival times according to some probability distributions. Considering the two options of choosing either the first regular or the first alternative connection, an interesting question bfor which probability distributions is one choice the better than the other. As there are many different probability distributions, it makes sense to limit ourselves to a certain class of distributions. In this section we discuss a way to solve this question for a uniform distribution and for an exponential distribution using binary search.

The expected arrival time of the regular connection

Let us assume we have access to a cumulative distribution function over the possible values of x defined as

$$f(z) = \Pr(x \leqslant z).$$

Let us also introduce a function $g : \mathbb{N}_0 \to [0, 1]$ such that $g(i)$ gives us the probability that the ith regular connection can pass the disrupted area, but the $(i-1)$th regular connection cannot yet. We calculate g using f in the following way:

$$g(i) = \begin{cases} f(\Delta S + y) & \text{if } i = 0 \\ f(\Delta S + iT_S + y) & \text{otherwise} \\ \quad - f(\Delta S + (i-1)T_S + y) & \end{cases}$$

Now the expected arrival time can be calculated using

$$E[\text{arr}] = \sum_{i=0}^{\infty} g(i)(\Delta S + iT_S + S).$$

If we want to evaluate this formula for a distribution with infinite support, we should terminate the calculation based on some small threshold ϵ and stop it when $f(\Delta S + iT_S + y) > 1 - \epsilon$.

Finding an indifference distribution

We first consider uniform distributions, where x is drawn uniformly from the range $[0, \omega]$ for a given $\omega > 0$. We will denote the expected arrival time for a given uniform distribution by $E_\omega[\text{arr}]$. We want to find an ω such that we are indifferent between both considered strategies, i.e., such that $E_\omega[\text{arr}] = \Delta D + D$. We can apply a binary search procedure in order to find an ω that brings us close enough to $\Delta D + D$. The ω that we find with this procedure can be used in the decision process as follows: If we believe it is likely that the real-life distribution behaves uniformly with range $[0, a]$ with $a < \omega$, we should prefer the regular connection over the alternative one. If we believe that $a > \omega$ we should prefer the alternative connection over the regular connection. We can apply a similar procedure if x is drawn from the exponential distribution with rate λ.

As long as we consider single parameter distributions for which the expected arrival $E[\text{arr}]$ is continuous when regarded as a function of the parameter of the distribution, and we can find lower and upper bounds for the parameter such that one bound gives $E[\text{arr}] < \Delta D + D$ and the other bound gives $E[\text{arr}] > \Delta D + D$, a binary search procedure can produce the value of the parameter for which the passenger will be indifferent between the regular and the alternative connections.

Although the indifference value can also be found analytically, we chose a binary search procedure since it is easy to implement and efficient enough to do computations within less than a second on almost any device. Note that also strategies that let the passenger wait some time and then depart are reasonable (and might minimize the expected arrival time dependent on the respective distribution). It will be part of our further research to include these strategies in the evaluation as well.

7.6 Implementation

As many passengers use smartphones during their journeys, we decided to implement the online and average case analysis using a platform that can be executed easily on these devices. Our implementation consists of roughly 250 lines of JavaScript, which can be executed by almost any web browser. This includes some rudimentary input checking and displaying results for the user. Our implementation is available at `http://computr.eu/pwp` and is able to calculate the results from Section 7.7 instantly on a modern smartphone.

Important for the performance of calculating ω and λ using binary search is the precision of the binary search in approximating $\Delta D + D$. In our implementation, the binary search terminates when $|E[arr] - (\Delta D + D)| < 0.001$. Furthermore, for the ϵ in evaluating $E[arr]$ for the exponential distribution we choose 0.001 as well. The expectation for the uniform distribution is evaluated with exact precision, as this distribution has finite support.

7.7 Examples

Table 7.1: Example instances with their parameters

Instance	ΔS	S	T_S	y	ΔD	D	T_D
A	1	15	15	0	5	30	20
B	1	15	15	5	5	30	20
C	5	10	10	5	2	60	15
D	1	10	60	5	10	60	5
E	10	10	60	5	1	60	5
Gd-Rtd	5	10	15	0	1	50	15
Ut-Rtd	5	30	15	20	2	85	30
Non-Convex	1	10	15	0	2	50	5

In order to evaluate our approach to the passenger waiting problem, we applied our calculation to a number of example instances, presented in Table 7.1. The results of our analysis are presented in Table 7.2. Although these results demonstrate our approach quite well, we encourage the readers to experiment with their own instances using our available web-based implementation from Section 7.6.

Table 7.2 can be read as follows: The top table contains the schedule definitions of the instances as discussed in Section 7.3. The bottom table contains the results: Our competitive analysis from Section 7.4 tells us that we should wait for the r^*th departure of the regular connection if we want to minimize the competitive ratio to

Table 7.2: Analysis results for the example instances from Table 7.1

Instance	r^*	$C_I^{r^*}$	ω	λ	$1/\lambda$
A	0	1.13	23	0.09	11.6
B	1	1.94	27.25	0.07	14.9
C	3	1.95	94.75	0.02	49.5
D	"any"	6.36	64	0.04	28.2
E	0	3.05	40.5	0.04	26.7
Gd - Rtd	2	1.80	67.3	0.03	34.1
Ut - Rtd	2	1.86	105.5	0.02	55.8
Non-Convex	6	1.58	68.5	0.03	34.8

$C_I^{r^*}$. The results of the average case analysis presented in Section 7.5 are given by the range ω of the uniform distribution and rate λ of the exponential distribution that make the passenger indifferent between taking the first detour train and waiting until the disruption is resolved. This means that if the passenger expects the disruption to vanish before the distributions listed in the table predict, he should wait until the disruption is resolved, while otherwise he should take the first detour train.

Let us first discuss instances A and B. With a regular travel time of 15 minutes and a detour of 30 minutes, the second regular connection will already arrive relatively close to the detour and it can thus be expected that waiting is not very beneficial. For instance A the competitive analysis reflects this with a choice for $r^* = 0$. Things change when we increase y from 0 to 5, as in instance B we get $r^* = 1$. The adversary then picks $x = 6$ and makes us end up in the second regular connection rather than in the alternative connection. However, the average case analysis yields an ω that is quite similar for instances A and B. This holds for $1/\lambda$ as well. As both values are not that large, this approach suggests to take the detour.

Instance C is an example where the competitive analysis suggests to wait longer. This is mostly because the alternative connection is quite lengthy compared to the regular connection. As a result, waiting for the next regular departure takes only a small amount of time compared to the increase in travel time of the detour. The large values of the uniform distribution with range $\omega \approx 95$ and the exponential distribution with mean $1/\lambda \approx 50$ suggest we should wait.

Instance D is interesting, as all values for r^* within the range considered yield the same competitive ratio. This is due to the adversary exploiting that $y > 0$ by letting the passenger take one regular connection later than the first one to pass through the disruption, by picking $x = 6$. Regardless of the waiting strategy, in the worst case the passenger will end up in the second departure of the regular connection.

If we swap ΔD and ΔS, we end up in instance E and we get a better competitive ratio if we always pick the first detour. The impact of the swap of ΔD and ΔS on the uniform range ω is quite large, while the exponential mean $1/\lambda$ barely changes. Therefore, interpreting the best strategy in this situation is quite difficult, but if we are risk-averse we can safely take the detour.

The examples presented in Figures 7.1 and 7.2 yield a competitive ratio of 1.80 and 1.86. In the Gd-Rtd scenario, it pays off to wait for 30 minutes to make sure that no early regular connection departs whose 10 minute trip length is much faster than the 60 minutes of the detour. For the Ut-Rtd scenario, the competitive ratio increases slightly, but the way in which it is attained is different. The best decision for the adversary is to release the disruption before the passenger stops waiting, while for the Gouda-Rotterdam case, it is better for the adversary to let the passenger take the detour.

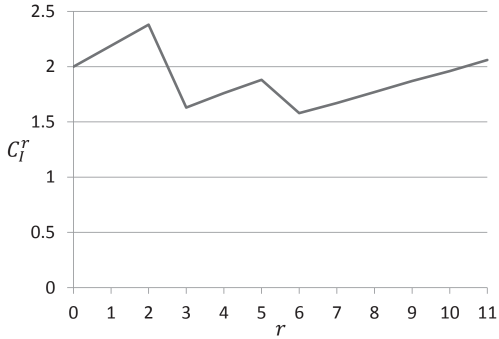


Figure 7.3: An example where the behavior of C_I^r as a function of r behaves non-convex

Finally, we have the non-convex scenario with a detour connection that is much more frequent than the regular one. This leads to non-convex behavior in the competitive ratio, as can be seen in Figure 7.3. Since there is no departure of the regular connection scheduled between the departures of the detour connection with $r = 1$ and $r = 3$, the detour connection with $r = 1$ dominates the detour connection with $r = 2$, as there is no opportunity to benefit from the resolution of the disruption in that interval. This example justifies that we indeed have to try a set of different values for r as suggested by Theorem 7.2.

To conclude, these example instances show that there is no strategy which is clearly superior in all cases, and the question which strategy is best depends both on the preferred quality measure and on the periodic timetables of the instance under consideration.

7.8 Conclusions and Future work

Considering our results with some examples instances, we can conclude that both approaches allow us to evaluate the trade-offs that occur within the passenger waiting problem. The average case analysis tells us for which probability distributions we should wait and for which we should take the detour, while the online analysis gives us a strategy with the best possible worst case performance.

We would like to extend our model to situations where the alternative connections have greater differences and may share edges within a network model. Moreover, we are working on a rigorous classification of instance classes for the passenger waiting problem based on their properties. Additionally, we think about good ways to communicate the output of our model to the end user. Finally, we consider how we can take preferences of the passengers, such as their valuation of waiting versus traveling and the riskiness of decisions, into account.

8

The Traveller's Route Choice Problem under Uncertainty – Dominance Strategies Under Uncertainty

Co-authors : Marie Schmidt, Leo Kroon, Anita Schöbel
This paper has been published in Operations Research.

8.1 Introduction

In practice, things often do not happen exactly as they were expected or planned to, as there will always be disturbances, inaccuracy of input data, and wrongly estimated or completely unknown parameters. Hence there is a need for optimization algorithms that are able to deal with such uncertainties. Such algorithms are developed in the field of *uncertain optimization* in which optimization problems with uncertain parameters are studied. The following different approaches have emerged. *Robust optimization* and *online optimization* aim at finding solutions which are best in the worst case. Different measures are used to evaluate the quality of these solutions, e.g., the objective value in the worst case, worst-case regret, or competitive ratio. *Stochastic optimization* additionally uses knowledge about underlying probability distributions to optimize the expected outcome or the probability of an outcome. Each of these approaches may lead to a different 'optimal' solution or strategy. In this chapter, we define the concept of (*weak*) *dominance* of strategies in uncertain situations. The set of non-dominated strategies can be considered a candidate set for the decision maker - we show that we only need to search within this set to find strategies which are optimal in the robust, online, or stochastic sense.

Our application is the *Traveler's Route Choice Problem* (*TRCP*) which was introduced in Chapter 7. We analyze dominance relations among strategies for this problem. The TRCP considers a traveler in a railway system with a periodic timetable who wants to travel from an origin to a destination and has chosen a route to do so. At some station he learns that his chosen route is blocked further on, and that it is uncertain at what point in time the blocked area will be opened again. Now the traveler has different options: He can board the train hoping that the disruption is over before the blocked area is reached. If this is not the case, he will have to wait directly in front of the blocked area until it is over. However, at his current station there is also another train service going to his destination. This other train service has a later arrival time at his destination than the service that was chosen initially, but it is not blocked and may hence be faster in the current situation. Now the question is whether the should traveler stick to the train service as planned or should he switch to a train along the detour route? Moreover, since we assume that both services are operated periodically, the traveler can also wait in the station and board one of the two trains later on. The TRCP is a relevant problem for railway operators; e.g., solution algorithms could be implemented in the route planning software to give good advice for travelers in situations as described above.

A situation as described does not only occur in railway systems, but also in other public transport systems and it can be generalized to other applications, e.g. to production systems, as will be explained. All these situations have in common that a decision on how to proceed has to be taken without exact knowledge on the duration of a disruption in the system. The decision to either hope that the preferred option will be in operation soon again or to take a detour is not trivial, since it depends on the frequencies and processing times of both alternatives and on the duration of the disruption, which is uncertain.

The contribution of this chapter is thus two-fold. On the one hand, we define the concept of (*weak*) *dominance* for general uncertain optimization or decision problems and show how it relates to quality measures for decision making and optimization under uncertainty from the literature.

On the other hand, we investigate dominance relations among the strategies for the TRCP. Based on the instance parameters, we are able to specify the set of non-dominated strategies. This often reduces the number of potentially optimal solutions considerably and will hence also reduce the computation time when solving this problem numerically.

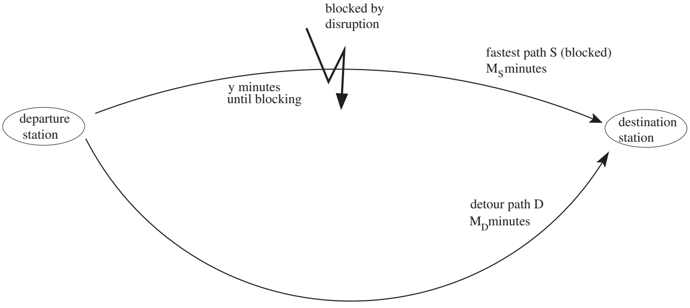


Figure 8.1: The fastest route S and the detour route D connecting the departure and the destination station.

The remainder of this chapter is structured as follows. Section 8.2 introduces a formal description of the TRCP, and shows that it has more applications than railway systems. Section 8.3 provides an overview of related literature. In Section 8.4 we describe the *strategies* that may be used for solving instances of the TRCP. Next, in Section 8.5 several optimization criteria from *robust, stochastic and online optimization* are defined, which can be used for evaluating strategies. In Section 8.6 we define the concept of (*weak*) *dominance* among strategies. Section 8.7 gives a complete description of the dominance relations among the strategies for the TRCP. Finally, Section 8.8 finishes the Chapter with conclusions and subjects for further research. The Appendix of this chapter, which is not included in this thesis but available as an appendix to the publication, contains additional graphics for the illustrative example in Section 8.5.2, and the proofs of the results in Sections 8.4, 8.6, and 8.7.

8.2 The Traveler's Route Choice Problem

8.2.1 Problem definition

Consider the following situation: A traveler in a railway system wants to travel by train from a departure station to a destination station. The fastest train service available at his wished departure time is a service along route S. The trains along this route are called *S-trains*. However, there is a disruption on this route such that it is blocked. The traveler has two options: Wait and hope that the disruption on the route S is over soon, or take the shortest available connection along a route with later arrival time D. The trains along the route D are called *D-trains*. See Figure 8.1 for an illustration of the situation.

Let (Δ_S, T_S, M_S) denote the schedule of the S-trains. The first S-train departs at time Δ_S, M_S denotes the (undisrupted) travel time of the S-trains, and T_S is the time period after which the schedule of the S-trains is repeated. That is, according to the timetable, the S-trains depart at the points in time Δ_S, $\Delta_S + T_S$, $\Delta_S + 2T_S, \ldots$, and arrive at the points in time $\Delta_S + M_S$, $\Delta_S + T_S + M_S$, $\Delta_S + 2T_S + M_S, \ldots$. Note that we have $T_S > \Delta_S$, since otherwise there would be one more S-train departing at time $0 < \Delta_S - T_S < \Delta_S$, hence the train departing at time Δ_S would not be the first one.

For the sake of simplicity we assume that the traveler learns about the disruption on the route S at time 0. Despite the disruption, the traveler can continue his trip on the route of the S-train until the blocked area. We denote by y the travel time from the origin until the blocked area is reached. If the disruption is over when the i-th S-train reaches the blocked area, that is if it is over before or at time $\Delta_S + iT_S + y$, then the train (and the traveler) can complete the journey undisruptedly. If the disruption is not over at that time, the train cannot continue and the traveler has to wait at the corresponding location until the disruption is over and he is picked up by the first S-train which runs again according to the regular timetable. This is common in real-world situations, where the timetable outside the disrupted area is operated according to a so-called *contingency plan*, which keeps the timetable there as much as possible the same as the regular timetable. This can be done by *short turning* of trains at both sides of the blocked area. Once the disruption is over, the regular timetable is resumed again.

Similarly, let (Δ_D, T_D, M_D) denote the schedule of the (fastest available) detour connection. Here Δ_D denotes the departure time of the first D-train after time 0, M_D is the travel time of the D-trains, and T_D is the time period after which the schedule of the D-trains is repeated. As above, we have $T_D > \Delta_D$. That is, the D-trains depart at the points in time Δ_D, $\Delta_D + T_D$, $\Delta_D + 2T_D, \ldots$, and arrive at $\Delta_D + M_D$, $\Delta_D + T_D + M_D$, $\Delta_D + 2T_D + M_D, \ldots$. All D-trains are operated.

We assume that $\Delta_S + M_S < \Delta_D + M_D$. That is, in case of no disruption, the traveler arrives earlier if he uses route S. Indeed, if $\Delta_S + M_S \geqslant \Delta_D + M_D$, then it is obvious that the traveler should take the first D-train anyway. Furthermore, we do *not* necessarily assume that $M_S < M_D$. We assume that all instance parameters are non-negative.

Let x denote the (unknown) point in time when the disruption is resolved. We analyze the following two situations in this chapter:

1. The maximum duration of the disruption is bounded by $x^{\max} < \infty$.
2. We have no prior information about the maximum duration of the disruption. In that case $x^{\max} = \infty$ and we use the notation $[a, x^{\max}] := [a, \infty)$ for intervals and $f(x^{\max}) := \lim_{x \to \infty} f(x)$ for functions of $x^{\max}$.

Thus an instance I of the TRCP with $I = ((\Delta_S, M_S, T_S), (\Delta_D, M_D, T_D), y, x^{\max})$ consists of

1. The schedules of the S-trains and the D-trains as well as the time y after which the disruption is reached on the route S.
2. A time bound $x^{\max} \in \mathbb{R}^+ \cup \{\infty\}$ which specifies the latest point in time when the disruption ends.

We finally introduce some notation which is convenient in the remainder of the Chapter. First, for each integer $r \in \mathbb{N}_0 \cup \{\infty\}$ we use

$$\mathrm{dep}_S(r) := \Delta_S + rT_S \quad \text{and} \quad \mathrm{dep}_D(r) := \Delta_D + rT_D$$

These time instants denote the time of the r-th departure of an S-train, or a D-train, respectively. Next, given a point in time x,

$$\mathrm{suc}_S(x) := \Delta_S + \max\left(0, \left\lceil \frac{x - \Delta_S}{T_S} \right\rceil\right) T_S$$

and

$$\mathrm{suc}_D(x) := \Delta_D + \max\left(0, \left\lceil \frac{x - \Delta_D}{T_D} \right\rceil\right) T_D$$

denote the time of the earliest departure of an S-train, or a D-train, respectively, after time x. We set $\mathrm{dep}_S(\infty) := \infty$, $\mathrm{dep}_D(\infty) := \infty$, $\mathrm{suc}_S(\infty) := \infty$, and $\mathrm{suc}_D(\infty) := \infty$.

8.2.2 Production planning: A second example for the TRCP setting

Railway systems are regularly confronted with disruptions. As a consequence, some of the scheduled services are canceled. This means that the TRCP is of great importance in railway systems. However, this is not the only application of the problem. One further application is exemplarily mentioned below.

Consider a company A producing some product. For the production, certain subparts are needed which are usually delivered by company B which is the fastest option for acquiring the required subparts. Company B accepts orders every day until 9:00 and usually delivers at 17:00.

Unfortunately, company B has some problems with a machine which is needed for the last two hours of the manufacturing process, but which currently cannot be used. It is uncertain when the machine's repair is finished such that the machine can be operated again.

However, there is another company, C, which is also able to produce the required subparts, but the production at company C takes 12 hours. Ordering the subparts at company C has to be done before 11:00.

Suppose it is 8:00 in the morning. The question now is: Should the order be given to company B or to company C? Or is it better to wait one day to see if company B is able to repair the machine until 15:00 when it is needed and then decide what to do?

This problem also fits in the framework of the TRCP by using an instance I with the following parameters: $I = ((\Delta_S = 1, M_S = 8, T_S = 24), (\Delta_D = 3, M_D = 12, T_D = 24), y = 15 - 9 = 6, x^{max} = \infty)$.

8.3 Literature Overview

Uncertainty in problem parameters can complicate decisions. This often concerns decisions which would be easy to make if all problem parameters were known, as for example in the TRCP. In the first paragraph of Section 8.3.1 we review literature on different models and quality measures for evaluating decisions under uncertainty. However, to compare strategies in this chapter, we do not use a specific quality measure for evaluating strategies, but instead make use of the more general concept of *(weak) dominance*. Various dominance concepts have been described in the literature. An overview can be found in the second paragraph of Section 8.3.1. Finally, in Section 8.3.2 we briefly discuss problems which show some similarity to the TRCP.

8.3.1 Decision making under uncertainty

Solution choice and optimization in uncertain situations Different models have evolved to explain people's choices in uncertain situations. Knight (1921), a pioneer in this field, distinguishes *risk*, a situation in which a probability distribution of an uncertain parameter is known, and *uncertainty*, where no probability distribution is given.

One of the most influential theories on decisions under risk is *expected utility theory*, see von Neumann and Morgenstern (1944). Every person is assumed to possess a utility function which assigns a value to each choice under each possible scenario and to choose the option which optimizes the *expected* utility. Other theories like *regret theory* (Loomes and Sugden, 1982; Bell, 1982; Fishburn, 1982) and *prospect theory* (Kahneman and Tversky, 1979) assume that decisions are not based only on the expected utility, but also on the expected regret in comparison to other options. This line of research is of a descriptive nature. It aims at modeling people's real-life behavior, not to provide a framework for making optimal decisions. The choice situations are often simple in the sense that there is only a limited number of options to choose from (often: only two) and that the outcome of a decision under all scenarios is known explicitly.

Decisions under uncertainty are also considered in the field of mathematical optimization. Research in this field aims at developing methods to choose the *best possible* solution for an optimization problem. Also here, the quality of a solution depends on both the solution and the scenario (which can hence be called a utility function in the above-used terminology). Besides, also feasibility of a solution may depend on the scenario. Similar to the above-described concepts, stochastic optimization often aims at optimizing the *expected* value (Birge and Louveaux, 1997). *Robust optimization* has different objectives: It searches for a solution which is optimal in the *worst-case* over all scenarios (Ben-Tal et al., 2009). Like above, besides absolute values, also relative values like *regret* or the *competitive ratio* from online optimization are considered in the literature (Kouvelis and Yu, 1997; Borodin and El-Yaniv, 1998). See also Section 8.5 for details on objectives used in stochastic, robust, and online optimization.

In these fields, the focus lies on developing optimization methods that are able to handle the typical problem structures arising from the modeling framework. Whether the (utility) function and the objective chosen is appropriate or whether it is the one an informed decision maker would base his/her decision upon is seldom discussed.

In the remainder of this chapter we will use the term *uncertainty* both for situations with and without probability distributions.

Dominance: Which solutions can be excluded? All concepts described in the previous section establish an order on the set of solutions. In the end the solution with the best value is selected.

However, different utility functions and different objectives lead to different orders of the solutions and hence also to different selected solutions. This can cause problems whenever the precise utility and/or objective is not known beforehand. In this case, it may be better to specify a set of solutions which contains all possible "reasonable" choices of solutions.

The concept of (first-order) stochastic dominance excludes all solutions which cannot be optimal with respect to the *expected value* of *any* increasing utility function. Narrowing down the set of considered utility functions leads to more restrictive dominance concepts (stochastic dominance of higher orders). See Levy (1992) for details. Stochastic (first order) dominance can alternatively be defined as follows: A solution $\mathcal{A}$ stochastically dominates a solution $\mathcal{B}$, if for each possible outcome α, the probability that $\mathcal{A}$ is better than α is higher than the probability that $\mathcal{B}$ is better than α. Similar dominance concepts have been described, e.g., in Perny and Spanjaard (2002), and Gabrel and Murat (2007) for uncertain situations without probability distributions. They correspond to first and second order stochastic dominance under the assumption of a uniform distribution on the set of possible scenarios.

However, whenever worst-case objectives are considered (as it is the case in robust and online optimization), stochastic dominance is not applicable, i.e. a solution may be optimal with respect to a worst-case objective although it is stochastically dominated. For this reason, in this chapter we are going to apply a dominance concept which excludes less solutions than stochastic dominance. It is motivated by Pareto-dominance in multi-objective optimization and sometimes referred to as state-wise or state-by-state dominance (see Section 8.6.1 for a definition, and, e.g. Ehrgott (2005) for the use in multi-objective optimization). We will show in Section 8.6.1 that any solution which is dominated (with respect to this concept) cannot be optimal with respect to any reasonable choice of objective and utility function.

This dominance concept is not new in optimization under uncertainty. It has been used, e.g. in Dias and Climaco (2000), Kouvelis and Sayin (2006), and Gast (2010) in the context of specific robust or stochastic optimization problems. A critical discussion of whether uncertain optimization problems can and should be regarded as multi-objective optimization problems in this manner is given in Hites et al. (2006). It is shown that some but not all properties can be transferred, that new insight is obtained, but that both notions must not be confused. A multi-objective perspective is also used in Gast (2010) in the context of supply chain management problems with uncertain input data. A unifying framework in which many robustness concepts are described by this multi-objective point of view is given in Klamroth et al. (2013) for finite scenario sets, and is under research for arbitrary uncertainty sets by Klamroth et al. (2015).

8.3.2 The TRCP and related problems

The TRCP The TRCP was first introduced in Chapter 7. In that Chapter, formulas for computing the competitive ratio for different strategies for the TRCP were given. Furthermore, assuming an underlying uniform or exponential distribution, it was analyzed whether departing immediately with the next D-train or waiting for the disruption to end leads to a lower expected arrival time, depending on the expected duration of the disruption. Based on these results, different strategies were evaluated on two cases from the Dutch railway network.

Uncertain shortest paths There exists a variety of uncertain versions of the shortest path problem, from the perspectives of stochastic, robust, and online optimization (Bar-Noy and Schieber, 1991; Nie and Wu, 2009; Yu and Yang, 1998; Westphal, 2008).

While some papers assume that a path has to be chosen *a priori*, others assume that it can be updated at any point in time (Bar-Noy and Schieber, 1991; Westphal, 2008). Two-stage models, where in the first stage an initial path is chosen which can be updated in a second stage, represent an intermediate concept between these two extremes (Büsing, 2012; Golovin et al., 2014).

A major difference between the described problems and the TRCP is that these problems are not schedule-based. Furthermore, in contrast to the majority of the models for uncertain shortest paths, in the TRCP the *length* of the disruption is uncertain.

Route choice in public transportation systems There is a huge number of publications about route choice in the *transit assignment* literature, see Liu et al. (2010) for an overview. The focus lies on modeling passenger flows as accurately as possible in regular traffic situations. In crowded situations, the focus is often on finding an equilibrium. The question of optimal route choice in case of disruptions is not addressed.

The problem of finding routes for (single) passengers is known as *timetable information*. Uncertain versions of this problem have been studied, e.g., in Dibbelt et al. (2013), Goerigk et al. (2013b), and Keyhani et al. (2014). In contrast to the TRCP which addresses the route choice in case of a major disruption, the mentioned papers focus on route choice in case of small delays.

Route choice in individual transportation A different problem which is studied in many different variants is the choice between two routes for individual travel. One of the routes is longer and/or tolled but generalized cost for traveling is deterministic. On the other route, travel time is uncertain. This problem is studied in the context of *learning* (i.e., the decision is *repeated* every day) in Chancelier et al. (2009). De Palma et al. (2012) study the value of different forms of information in this situation. Lam and Small (2001), and Avineri and Prashker (2003) quantify a traveler's sensitivity to uncertainty in such a setting.

8.4 Strategies for the uncertain Traveler's Route Choice Problem

8.4.1 The deterministic version

Before we analyze the uncertain case which was described as the TRCP in Section 8.2.1, we consider the deterministic version of the Traveler's Route Choice Problem. This is useful since we want to compare the solutions for instances of the TRCP with the solutions that would have been obtained if the end time of the disruption had been known in advance.

The deterministic version of the Traveler's Route Choice Problem is the same problem as the TRCP, but now also the end time of the disruption x is given in advance. In this problem the question is again: Which connection should the traveler choose, such that his arrival time at his destination is earliest? Fortunately, this question can be answered easily and optimally by comparing the arrival times of the S-trains and the D-trains:

Lemma 8.1. *For a given instance* I *of the deterministic Traveler's Route Choice Problem with* $I = ((\Delta_S, M_S, T_S), (\Delta_D, M_D, T_D), y, x)$, *the minimal arrival time* $z^*(x)$ *is given by*

$$z^*(x) = \min\{\mathrm{suc}_S(x - y) + M_S, \Delta_D + M_D\}.$$

The proof of this result is given in the online appendix of the publication of this chapter.

8.4.2 The uncertain version

In the TRCP, the traveler only knows that the route S is disrupted, however, he does not know the end of the disruption before the disruption is actually over at time x.

As long as the traveler has not taken one of the trains at any time before time t, his options at time t are:

- Remain waiting at the departure station (W).
- Take an S-train (S): This is only possible if $t = \mathrm{dep}_S(r) = \Delta_S + rT_S$ for some $r \in \mathbb{N}_0$.
- Take a D-train (D): This is only possible if $t = \mathrm{dep}_D(r) = \Delta_D + rT_D$ for some $r \in \mathbb{N}_0$.

The described options can be represented in a decision tree as shown in Figure 8.2 where each directed path from the root node (the leftmost node) to a leaf of the tree represents a strategy of the traveler. We assume that, as soon as the traveler has boarded a train, he cannot withdraw this decision. That is, he cannot go back to the departure station, not even if $y = 0$.

The decision of the traveler about how to proceed with his journey can be based on a so-called *strategy*. A strategy consists of two stages, and describes:

- How long the traveler intends to wait at the departure station until boarding one of the trains (this is the first stage of the strategy).
- How the traveler reacts if the disruption is over *while he is still waiting at the departure station* (this is the second stage of the strategy).

Note that the traveler may not have to deal with the second stage, since he took one of the trains before the disruption was over. Then he cannot leave the train again, but is bound to his first stage decision.

The first stage

In the *first stage* of the strategy, the traveler chooses one of the leaves of the decision tree in Figure 8.2, corresponding to the train he intends to take. He may choose to take an S-train at time $\mathrm{dep}_S(r) = \Delta_S + rT_S$.

In this case the traveler arrives at time: $\begin{cases} \mathrm{dep}_S(r) + M_S & \text{if } x \leqslant \mathrm{dep}_S(r) + y \\ \mathrm{suc}_S(x - y) + M_S & \text{otherwise} \end{cases}$ (8.1)

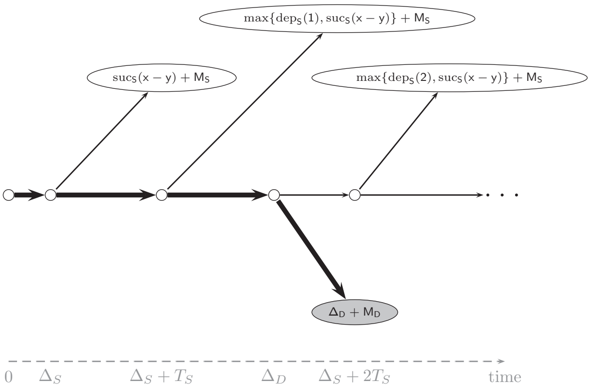


Figure 8.2: The first stage path corresponding to the strategy "take the first D-train" (later called $\mathcal{A}_{0,\mathrm{D}}$).

Indeed, either the disruption is over when the S-train arrives at the disrupted area (i.e., $x \leqslant \mathrm{dep}_\mathrm{S}(r) + y$), or the traveler has to wait at the point where the disruption takes place and can continue his journey with the first regular S-train once the disruption is over (see also Lemma 8.1). He may also choose to take a D-train at time $\mathrm{dep}_\mathrm{D}(r) = \Delta_\mathrm{D} + rT_\mathrm{D}$.

$$\text{In this case the traveler arrives at time:} \quad \mathrm{dep}_\mathrm{D}(r) + M_\mathrm{D}. \tag{8.2}$$

In Figure 8.2, these arrival times are depicted in the corresponding leaves of the decision tree. Note that, for a given instance of the TRCP, the arrival times of the D-trains are constant, and the arrival times of the S-trains depend only on the (unknown) end time of the disruption x. Thus as long as the end of the disruption x is unknown, there is no need nor use to change the first stage strategy.

The second stage

If the disruption is over while the traveler is still waiting at the departure station, it is reasonable to adapt his earlier chosen first stage strategy, for example, by taking the next S-train instead of the train chosen in the first stage. Indeed, at that moment the arrival times of all strategies can be evaluated, as is shown in Figure 8.3. Hence, the train with the earliest arrival time can be selected. This is the *second stage* of the strategy.

Since the end of the disruption x is revealed at time $t = x$, this implies that a reasonably acting traveler chooses to act as follows:

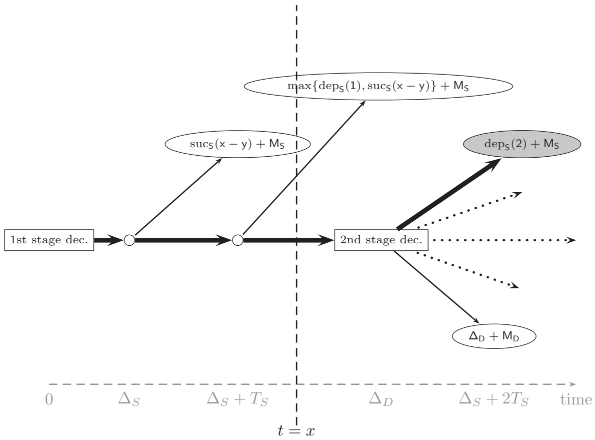


Figure 8.3: The second stage path corresponding to the strategy "take the first D-train" (later called $\mathcal{A}_{0,D}$) for a realization of x and the case $\mathrm{dep}_S(2) + M_S < \Delta_D + M_D$.

- If $\mathrm{suc}_S(x) + M_S < \mathrm{suc}_D(x) + M_D$, then he takes the next available S-train at time $\mathrm{suc}_S(x)$ and he arrives at time $\mathrm{suc}_S(x) + M_S$.
- If $\mathrm{suc}_S(x) + M_S > \mathrm{suc}_D(x) + M_D$, then he takes the next available D-train at time $\mathrm{suc}_D(x)$, and he arrives at time $\mathrm{suc}_D(x) + M_D$.

$$\text{Thus in both cases the traveler arrives at time:} \quad \min\{\mathrm{suc}_S(x) + M_S, \mathrm{suc}_D(x) + M_D\} \tag{8.3}$$

Without loss of generality, we assume that the traveler takes the S-train if $\mathrm{suc}_S(x) + M_S = \mathrm{suc}_D(x) + M_D$. Note that, if the traveler is still waiting at the departure station at a point in time t with $\mathrm{suc}_D(t) + M_D < \mathrm{suc}_S(t) + M_S$, then he should take the D-train anyway, whether the disruption is over or not.

Combining the first and second stage

As was described before, if the disruption ends at time x while the traveler is still waiting at the departure station, he takes the *uniquely defined* second-stage strategy $\mathcal{A}^*(x)$. Hence, the *practical* strategies for the TRCP can be described as follows:

- The strategies $\mathcal{A}_{r,S}$, for $r = 0, \ldots, \infty$ connected to the r-th departure of an S-train:

  - Wait at the departure station not longer than until $\min\{x, \mathrm{dep}_S(r)\}$, i.e., until the r-th S-train departs or until the disruption is over, whatever happens first.
  - If $x \leqslant \mathrm{dep}_S(r)$ use the second-stage strategy $\mathcal{A}^*(x)$.
  - If $\mathrm{dep}_S(r) < x$, take the r-th S-train.
- The strategies $\mathcal{A}_{r,D}$, for $r = 0, \ldots, \infty$ connected to the r-th departure of a D-train:
  - Wait at the departure station not longer than until $\min\{x, \mathrm{dep}_D(r)\}$, i.e., until the r-th D-train departs or until the disruption is over, whatever happens first.
  - If $x \leqslant \mathrm{dep}_D(r)$ use the second-stage strategy $\mathcal{A}^*(x)$.
  - If $\mathrm{dep}_D(r) < x$, take the r-th D-train.

Note that the above strategies are the strategies that are not sorted out as being impractical *at first sight*. In Section 8.7 on dominance among strategies for the TRCP we show that some of the above strategies are clearly superior to others.

If $\mathcal{A}$ is one of the above defined strategies for solving instances of the TRCP, then we will use the notation $z(\mathcal{A}, x)$ for the arrival time that is realized by applying strategy $\mathcal{A}$, if the end time of the disruption turns out to be x. Based on these descriptions, the arrival times realized by the strategies $\mathcal{A}_{r,S}$ and $\mathcal{A}_{r,D}$ can be computed as follows:

$$z(\mathcal{A}_{r,S}, x) = \begin{cases} \min\{\mathrm{suc}_S(x) + M_S, \mathrm{suc}_D(x) + M_D\} & \text{if } x \leqslant \mathrm{dep}_S(r), & \text{see (8.3)} \\ \mathrm{dep}_S(r) + M_S & \text{if } \mathrm{dep}_S(r) < x & \\ & \text{and } x \leqslant \mathrm{dep}_S(r) + y, & \text{see (8.1)} \\ \mathrm{suc}_S(x - y) + M_S & \text{otherwise,} & \text{see (8.1)} \end{cases}$$

$$z(\mathcal{A}_{r,D}, x) = \begin{cases} \min\{\mathrm{suc}_S(x) + M_S, \mathrm{suc}_D(x) + M_D\} & \text{if } x \leqslant \mathrm{dep}_D(r), & \text{see (8.3)} \\ \mathrm{dep}_D(r) + M_D & \text{otherwise,} & \text{see (8.2)} \end{cases}$$

8.5 Evaluating strategies under uncertainty

8.5.1 Quality measures for optimization under uncertainty

There are many different ways of measuring the quality of a strategy under uncertainty. Consider an optimization problem

$$\min_{\mathcal{A} \in \mathcal{S}} z(\mathcal{A}, x),$$

where the objective function z depends on a strategy $\mathcal{A}$ and on an uncertain parameter x from an *uncertainty set* $\mathcal{X}$. In the following, we briefly describe some common quality measures used for evaluation under uncertainty. Note that most literature on robust and stochastic optimization deals with *solutions*, however, we transfer these concepts to *strategies* in the following description.

Robust optimization aims at the minimization of some objective function in the *worst case*. In *strict robustness*, the worst-case *absolute* objective value

$$g^{\text{wc-abs}}(\mathcal{A}) := \sup_{x \in \mathcal{X}} z(\mathcal{A}, x)$$

is used as an objective function, see e.g. Ben-Tal et al. (2009). Other robustness approaches do not evaluate the absolute objective value over all scenarios x, but compare the realized objective value of a strategy $\mathcal{A}$ to the objective value of the best possible strategy that could have been obtained under prior knowledge of scenario x, denoted as $z^*(x)$. We call the difference

$$z^{\text{rg}}(\mathcal{A}, x) := z(\mathcal{A}, x) - z^*(x)$$

the *absolute regret* of the strategy $\mathcal{A}$ in scenario x. Similarly, the ratio

$$z^{\text{comp}}(\mathcal{A}, x) := \frac{z(\mathcal{A}, x)}{z^*(x)}$$

is called the *competitive ratio* (or *relative regret*) of the strategy $\mathcal{A}$ in scenario x. Minimizing the *worst-case regret*

$$g^{\text{wc-rg}}(\mathcal{A}) := \sup_{x \in \mathcal{X}} z(\mathcal{A}, x) - z^*(x)$$

and the *worst-case competitive ratio*

$$g^{\text{wc-comp}}(\mathcal{A}) := \sup_{x \in \mathcal{X}} \frac{z(\mathcal{A}, x)}{z^*(x)}$$

are common objective functions in robust optimization, see e.g. Kouvelis and Yu (1997). In online optimization, the competitive ratio is the most common quality measure, see e.g. Borodin and El-Yaniv (1998).

If a probability distribution with probability density function p on the uncertainty set $\mathcal{X}$ is known, the conservative approach of robust optimization can be weakened by excluding unlikely scenarios with high objective values. For example, for any $\gamma \in [0, 1]$ and $f \in \{z, z^{\text{rg}}, z^{\text{comp}}\}$ we define by

$$g^{\text{reach-}\gamma}(\mathcal{A}) := \min\{\alpha : \int_{I(\mathcal{A}, \alpha)} p(x)dx \geqslant \gamma\},$$

with

$$I(\mathcal{A}, \alpha) := \{x : f(\mathcal{A}, x) \leqslant \alpha\}, \tag{8.4}$$

the minimal value of f which can be guaranteed with probability γ. If we set $\gamma := 1$, we obtain the above-described problems of minimizing the worst-case absolute value, regret or competitive ratio from robust and online optimization. An example for this objective function where $\gamma < 1$ can be found, e.g. in Daskin et al. (1997) and Gao (2011).

A different objective which is often used in stochastic optimization is the *expected value*

$$g^{\exp}(\mathcal{A}) := \int_{\mathcal{X}} f(\mathcal{A}, x) p(x) dx$$

(where $p : \mathcal{X} \to [0, 1]$ is the probability distribution on the uncertainty set $\mathcal{X}$), which, as before, can be evaluated for $f \in \{z, z^{rg}, z^{comp}\}$, see e.g. Birge and Louveaux (1997).

If the traveler in the TRCP has an important appointment at his destination at a certain time, then he would choose a strategy with *highest probability* to arrive on time. Likewise, the probability to achieve a certain regret value or competitive ratio can be maximized. Let $I(\mathcal{A}, \alpha)$ be defined as in (8.4). We define

$$g^{\text{prob-}\alpha}(\mathcal{A}) := \int_{I(\mathcal{A},\alpha)} p(x) dx$$

as the probability that $f(\mathcal{A}, x)$ stays below the value α for $f \in \{z, z^{rg}, z^{comp}\}$.

This function $g^{\text{prob-}\alpha}(\mathcal{A})$ is sometimes called the *reliability* of $\mathcal{A}$ (Nie and Wu, 2009; Pan et al., 2013). Often, a lower bound on the reliability, a so-called *chance constraint* $g^{\text{prob-}\alpha}(\mathcal{A}) \geqslant \gamma$ is imposed, see e.g. Valdebenito and Schuëller (2010) and Birge and Louveaux (1997). Approaches that find the most reliable strategy are less common, see e.g. Nie and Wu (2009); Gao (2011); Pan et al. (2013).

As a last example for a quality measure, some papers, see e.g. Sigal et al. (1980), also investigate how to find a strategy which has the highest probability of being optimal, i.e., they take

$$g^{\text{opt}}(\mathcal{A}) := \int_{I(\mathcal{A})^{\text{opt}}} p(x) dx$$

with $I(\mathcal{A})^{\text{opt}} := \{x : f(\mathcal{A}, x) \leqslant f(\mathcal{A}', x) \text{ for all } \mathcal{A}'\}$, as an evaluation criterion. Note that for this measure it is irrelevant whether we choose f as the absolute value, the regret or the competitive ratio.

Summarizing, which strategy is considered as optimal depends on:

1. The way of measuring the objective value for a given $\mathcal{A}$ and a given scenario x:
   - Absolute value

- Absolute regret
- Competitive ratio (or relative regret)

2. The *utility function*, i.e., the way of aggregating over the uncertainty set:
   - Minimize: Objective value that can be guaranteed with probability at least γ (with robust and online optimization ($\gamma = 1$) as special cases)
   - Minimize: Average value
   - Maximize: Probability of reaching a predefined value α
   - Maximize: Probability of having found an optimal solution
3. The choice of the probability distribution.

Hence, while in the deterministic case the definition of an *optimal solution* is straightforward, this is not true in the uncertain case. In this chapter we compare different strategies $\mathcal{A}$. The uncertain parameter in this case is the time x at which the disruption vanishes. In the next section we indeed see that different quality measures lead to different "optimal" solutions.

8.5.2 An Illustrative Example

Consider an instance of the TRCP where the timetable of the S-train is defined by $\Delta_S = 0$, $T_S = 2$, $M_S = 2$, and the timetable of the D-train is defined by $\Delta_D = 0$, $T_D = 5$, $M_D = 12$. We furthermore assume that the blocked area on the route of the S-train is reached after $y = 1$, and that the length of the disruption is unbounded with an expected duration of 11. All times are specified in minutes.

Table 8.1 illustrates that in this example, optimality of a strategy strongly depends on the quality measure. To obtain better insights on how the strategies perform for different values of x it is advisable to consult the appendix. Table 8.1 shows the different quality measures for the strategies $\mathcal{A}_{r,D}$ and $\mathcal{A}_{r,S}$ for $r = 0, 1, 2$.

The first subtable shows worst-case values for the (absolute) travel time, the regret and the competitive ratio obtained in the deterministic version of the problem. Optimizing according to these criteria leads to different results. If the traveler wants to be at the destination as early as possible in the worst-case, he should follow strategy $\mathcal{A}_{0,D}$ because this is the smallest value in the first row. If a relative quality measure, i.e., the regret or the competitive ratio, is used, the situation is different. In this example, for any D-train strategy $\mathcal{A}_{r,D}$, the regret is maximal if the disruption is resolved just after the traveler took the D-train, and is then basically equal to the difference in travel times $M_D - M_S = 10$. However, the second (and any further) D-train strategy has a slight advantage, due to the fact that the frequency of the D-train is not a multiple of that of the S-train. For the competitive ratio the best strategy is to wait for the third D-train. Since in the worst case the disruption may never be resolved, the strategies

	take D-train			take S-train		
	$\mathcal{A}_{0,D}$	$\mathcal{A}_{1,D}$	$\mathcal{A}_{2,D}$	$\mathcal{A}_{0,S}$	$\mathcal{A}_{1,S}$	$\mathcal{A}_{2,S}$
1. Worst-case value (guaranteed with probability 1)						
absolute value	**12**	17	22	∞	∞	∞
regret	10	**9**	10	∞	∞	∞
ratio	6	2.13	**2**	∞	∞	∞
2. Value guaranteed with given probability γ for exponential/Poisson distribution with mean $\lambda = 11$						
absolute value (for $\gamma = 0.5$)	12/**12**	17/17	**10**/22	**10**/**12**	**10**/**12**	**10**/**12**
absolute value (for $\gamma = 0.9$)	**12**/**12**	17/17	22/22	28/16	28/16	28/16
regret (for $\gamma = 0.5$)	2/**0**	5/5	2/10	**0**/**0**	**0**/**0**	2/**0**
regret (for $\gamma = 0.9$)	**8/4**	9/9	10/10	16/**4**	16/**4**	16/**4**
ratio (for $\gamma = 0.5$)	1.2/**1**	1.42/1.42	1.5/1.83	**1**/**1**	**1**/**1**	1.17/**1**
ratio (for $\gamma = 0.9$)	3/1.5	2.23/2.23	**1.83**/1.83	2.33/**1.33**	2.33/**1.33**	2.33/**1.33**
3. Expected value for the exponential/Poisson distribution with mean $\lambda = 11$						
absolute value	**12/12**	12.77/16.63	13.21/16.75	12.97/12.50	13.14/12.50	13.29/12.51
regret	**3.44/1.05**	4.21/5.68	4.65/5.80	4.41/1.55	4.59/1.55	4.73/1.56
ratio	1.93/1.14	1.51/1.54	1.50/1.50	**1.37**/**1.13**	1.45/1.13	1.49/1.13
4. Probability of reaching value (in %) for the exponential/Poisson distribution with mean $\lambda = 11$						
absolute $\leqslant 10$	0/<1	37/4	52/23	**56/34**	**56/34**	**56/34**
absolute $\leqslant 12$	**100**/**100**	37/4	60/46	63/58	63/58	63/58
regret $\leqslant 2$	53/**86**	37/4	60/46	**70**/78	**70**/78	69/78
regret $\leqslant 5$	63/**96**	**81**/70	60/46	74/91	74/91	74/91
competitive ratio $\leqslant 1.5$	63/**96**	**72**/70	51/46	63/58	55/58	47/58
competitive ratio $\leqslant 2$	76/> 99	90/89	**100**/**100**	88/> 99	88/> 99	88/> 99
5. Probability of being optimal (in %) for the exponential/Poisson distribution with mean $\lambda = 11$						
any measure	44/**66**	15/1	28/26	**63**/58	55/58	50/47

Table 8.1: Objective values for different quality measures for the strategies $\mathcal{A}_{r,P}$ with $r \in \{0, 1, 2\}$ and $P \in \{S, D\}$. In each row, the optimal value(s) are shown in **bold**.

$\mathcal{A}_{r,S}$ have infinite objective value and should not be used when planning for the worst-case.

When looking at probabilities instead of worst cases, as we do in subtables 2-5, the situation is different. Since the expected duration of the disruption is only 11 minutes, which is lower than the driving time of the detour train (12 minutes), with respect to some of the quality measures it turns out to be better to follow an S-train strategy than to follow a D-train strategy.

For example, we can see in the third line of the second subtable that if the traveler follows strategy $\mathcal{A}_{0,S}$, he has a 50% chance to a regret of 0. That means that with (at least) 50% probability it is optimal to take the first S-train, if the disruption length is distributed exponentially. The D-train strategies, on the contrary, will yield a higher regret with probability 50%.

Of course not only the expected value of the disruption length matters, but also the assumption on the underlying distribution. For example, if the traveler wants to optimize the probability of traveling at most 2 minutes longer than necessary, he should follow strategy $\mathcal{A}_{0,S}$ or $\mathcal{A}_{1,S}$ if he assumes an underlying exponential distribution, but under the assumption of a Poisson distribution he should follow strategy $\mathcal{A}_{0,D}$ (line 3 of subtable 4).

A particular case can be observed in the first line of the fourth subtable: Consider a traveler who has an important appointment at his destination and has to arrive there within the next 10 minutes. Since the travel time of the D-train is 12 minutes, the passenger's only chance for an arrival within the next 10 minutes is that the disruption is resolved fast and S-trains are operational soon again. That is the reason why the passenger should apply one of the strategies $\mathcal{A}_{r,S}$ to maximize the probability of timely arrival - and why D-train strategies which imply waiting longer for the disruption to vanish have a much better objective value than $\mathcal{A}_{0,D}$.

8.6 Comparing strategies under uncertainty: (weak) dominance

8.6.1 The concept of (weak) dominance

In the example in Section 8.5.2, no strategy is *optimal* with respect to all criteria - *optimality*, in fact, depends on the quality measure used for evaluating possible strategies, i.e., on how to handle the uncertain length of the disruption. However, we observe that some strategies in our example are inferior to others: e.g. $\mathcal{A}_{1,S}$ is not better than $\mathcal{A}_{0,S}$ in any of the listed cases - but it is worse in some of them. This idea is captured by the concept of (weak) dominance which we define in the following. Roughly speaking, a strategy (weakly) dominates another strategy if the result of the first one is at least as good as the result of the second one for every possible scenario $x \in \mathcal{X}$. Thus, when looking for an optimal strategy, we can ignore the second one.

Definitions 5 and 6 formally describe the concepts of *weak dominance* and *dominance* of one strategy over another strategy.

Definition 5. *Let* I *be an instance of an optimization problem under uncertainty. A strategy* $\mathcal{A}$ weakly dominates *another strategy* $\mathcal{A}'$ *on the instance* I *if* $z(\mathcal{A}, x) \leqslant z(\mathcal{A}', x)$ *for all* $x \in \mathcal{X}$.

Definition 6. *Let* I *be an instance of an optimization problem under uncertainty. A strategy* $\mathcal{A}$ dominates *another strategy* $\mathcal{A}'$ *on the instance* I *if* $\mathcal{A}$ weakly dominates $\mathcal{A}'$ *and if, additionally,* $z(\mathcal{A}, x) < z(\mathcal{A}', x)$ *for at least one* $x \in \mathcal{X}$.

The fact that a rationally-acting decision maker would not choose a dominated solution is widely accepted in the literature. In Lemma 8.2 we justify this assumption by showing that none of the concepts for measuring solution quality under uncertainty with objective function g described in Section 8.5.1 would lead to a dominated solution. A formal proof is given in the appendix. This is a generalization of the results in Gabrel and Murat (2007), Klamroth et al. (2013), and Iancu and Trichakis (2014) where this result was proven for some specific concepts for measuring solution quality.

Lemma 8.2. *If strategy* $\mathcal{A}$ *weakly dominates strategy* $\mathcal{A}'$*, then* $g(\mathcal{A}) \leqslant g(\mathcal{A}')$.

We conclude that, although we cannot clearly say which strategy is *best* under uncertainty without knowing a decision maker's exact way of evaluating solution quality, we can exclude all dominated strategies, since we have seen that, if a strategy $\mathcal{A}$ weakly dominates a strategy $\mathcal{A}'$, then $\mathcal{A}$ is at least as good as $\mathcal{A}'$ with respect to *any reasonable objective function*. Furthermore, note that, since stochastic dominance implies dominance as it is defined here, sorting out the dominated solutions will not remove any stochastically non-dominated solutions.

8.6.2 Non-dominated strategies in the illustrative example

We now illustrate the concept of dominated and non-dominated strategies in the example from Section 8.5.2 with $\Delta_S = 0$, $T_S = 2$, $M_S = 2$, $\Delta_D = 0$, $T_D = 5$, $M_D = 12$ and $y = 1$. In the reasoning below we make use of the fact that $\mathrm{suc}_S(x) + 2 < \mathrm{suc}_D(x) + 12$ for all x. Thus, if the disruption is over at time x and the passenger is still waiting at that time, then it is best to take the first S-train, independent of the value of x.

We first consider the D-train strategies. We see in Table 8.1 that the strategies $\mathcal{A}_{0,D}$, $\mathcal{A}_{1,D}$, $\mathcal{A}_{2,D}$ are optimal for certain choices of the objective. According to Lemma 8.2, we can conclude that they are non-dominated. However, we can even show that *all* D-train strategies $\mathcal{A}_{r,D}$ are not weakly dominated in this example:

1. First we show that there is no r' such that $\mathcal{A}_{r',D}$ dominates $\mathcal{A}_{r,D}$. To this purpose, for a given r' we assume a disruption length of $x = r'T_D + 1$. We obtain a travel time of
$$z(\mathcal{A}_{r',D}, r'T_D + 1) = r'T_D + M_D = 5r' + 12$$
for strategy $\mathcal{A}_{r',D}$. We now compare this to the travel time of strategy $\mathcal{A}_{r,D}$.
   - If $r' < r$, then $\mathcal{A}_{r,D}$ yields a travel time of
   $$z(\mathcal{A}_{r,D}, r'T_D+1) = \mathrm{suc}_S(r'T_D+1)+M_S \leqslant r'T_D+1+(T_S-1)+M_S = 5r'+4.$$
   - If $r' > r$, then $\mathcal{A}_{r,D}$ yields a travel time of
   $$z(\mathcal{A}_{r,D}, r'T_D + 1) = rT_D + M_D = 5r + 12.$$

   Hence in both cases, for a disruption length of $x = r'T_D + 1$, strategy $\mathcal{A}_{r,D}$ is better than strategy $\mathcal{A}_{r',D}$. According to Definition 5 it follows that $\mathcal{A}_{r',D}$ does not weakly dominate $\mathcal{A}_{r,D}$.

2. Now we show that there is no r' such that $\mathcal{A}_{r',S}$ dominates $\mathcal{A}_{r,D}$. To this purpose, for a given r' we assume a disruption length of $x = rT_D + M_D$. .
   - If $rT_D + M_D > r'T_S$, then for $\mathcal{A}_{r',S}$ we obtain a travel time of
   $$z(\mathcal{A}_{r',S}, rT_D + M_D) = \mathrm{suc}_S(5r + 12 - 1) + 2 \geqslant 5r + 13.$$
   - If $rT_D + M_D \leqslant r'T_D$, then for $\mathcal{A}_{r',S}$ we obtain a travel time of
   $$z(\mathcal{A}_{r',S}, rT_D + M_D) = \mathrm{suc}_S(5r + 12) + 2 \geqslant 5r + 14.$$

   We now compare this to the travel time of strategy $\mathcal{A}_{r,D}$ which is
   $$z(\mathcal{A}_{r,D}, rT_D + M_D) = rT_D + M_D = 5r + 12.$$
   Hence for a disruption length of $x = rT_D + M_D$, strategy $\mathcal{A}_{r,D}$ is better than strategy $\mathcal{A}_{r',S}$. According to Definition 5 it follows that $\mathcal{A}_{r',S}$ does not weakly dominate $\mathcal{A}_{r,D}$.

We can hence conclude that all D-train strategies are not weakly dominated in this example.

We now consider the S-train strategies. In Table 8.1 we see that $\mathcal{A}_{0,S}$ is optimal for several objective functions, and that for some it is the only optimal solution (at least among the strategies shown in the table). We conclude according to Lemma 8.2 that $\mathcal{A}_{0,S}$ is not weakly dominated. Strategies $\mathcal{A}_{r,S}$ for $r \in \mathbb{N}$ however, are dominated by

$\mathcal{A}_{0,S}$ in this example. We prove this by showing that for any $x \in [0, \infty)$, $z(\mathcal{A}_{0,S}, x)$ is at least as good as $z(\mathcal{A}_{r,S}, x)$ and for some x it is strictly better:

- if $x \leqslant 1$, then $z(\mathcal{A}_{r,S}, x) = 2 + 2 > 2 = z(\mathcal{A}_{0,S}, x)$
- if $x \in (1, 2r]$, then $z(\mathcal{A}_{r,S}, x) = \mathrm{suc}_S(x) + 2 \geqslant \mathrm{suc}_S(x-1) + 2 = z(\mathcal{A}_{0,S}, x)$
- if $x \in (2r, 2r+1]$, then $z(\mathcal{A}_{r,S}, x) = 2r + 2 = z(\mathcal{A}_{0,S}, x)$
- if $x > 2r + 1$ then $z(\mathcal{A}_{r,S}, x) = \mathrm{suc}_S(x-1) + 2 = z(\mathcal{A}_{0,S}, x)$.

Summarizing, in the illustrative example from Section 8.5.2, all D-train strategies are non-dominated, but only one of the S-train strategies is non-dominated, namely $\mathcal{A}_{0,S}$.

8.7 Results on dominance for the TRCP

Lemma 8.2 shows that for finding an optimal strategy with respect to any of the quality measures described in Section 8.5.1 we only have to investigate non-dominated strategies. This motivates our research on dominance in this section.

As before, the uncertainty set $\mathcal{X}$ equals the interval $[0, x^{max}]$, where $x^{max} \in \mathbb{R}^+ \cup \{\infty\}$. For the sake of a more compact representation, all proofs of the theorems and lemmas in this section are given in the appendix.

In this section, we describe the dominance relations among the D-train strategies $\mathcal{A}_{r,D}$, among the S-train strategies $\mathcal{A}_{r,S}$, and between the $\mathcal{A}_{r,D}$ and the $\mathcal{A}_{r,S}$ strategies. We start with the following intuitive result, which allows us to exclude some trivial cases.

Lemma 8.3.

- *If* $\Delta_D + M_D \leqslant \Delta_S + M_S$, *then the D-train strategy* $\mathcal{A}_{0,D}$ *weakly dominates all other strategies.*
- *If* $\mathrm{suc}_S(x^{\max} - y) + M_S \leqslant \Delta_D + M_D$, *then the S-train strategy* $\mathcal{A}_{0,S}$ *weakly dominates all other strategies.*

Based on the result of Lemma 8.3, we consider in this section only instances I of the TRCP satisfying the following condition:

$$\Delta_S + M_S < \Delta_D + M_D < \mathrm{suc}_S(x^{\max} - y) + M_S. \tag{8.5}$$

The remainder of this section is structured as follows. In Section 8.7.1 we show the intuitive results that the D-train strategy $\mathcal{A}_{0,D}$ is not weakly dominated, and that in most cases the S-train strategy $\mathcal{A}_{0,S}$ is not weakly dominated either. Section 8.7.2 gives necessary and sufficient conditions for a D-train strategy to be (weakly) dominated by another D-train strategy. In Section 8.7.3, (weak) dominance among S-train strategies is studied for the case $M_S < M_D$. The main result here is that for $y > 0$ all S-train strategies $\mathcal{A}_{r,S}$ with $r \in \mathbb{N}$ are dominated. The case $M_D \leqslant M_S$ is described in Section 8.7.4. In this case the only non-dominated strategies are $\mathcal{A}_{0,S}$ (if $y > 0$) and $\mathcal{A}_{0,D}$. In Section 8.7.5, we consider the case $y = 0$. In this case it is better to wait until the disruption is over than to commit oneself to any S-train. Finally, the results that have been derived in this section are summarized in a table in Section 8.7.6.

8.7.1 Dominance of $\mathcal{A}_{0,S}$ and $\mathcal{A}_{0,D}$

In this section we describe the rather intuitive results that in most cases the S-train strategy $\mathcal{A}_{0,S}$ and the D-train strategy $\mathcal{A}_{0,D}$ are not weakly dominated. Note that this does *not necessarily* mean that in most cases $\mathcal{A}_{0,S}$ and $\mathcal{A}_{0,D}$ are the best strategies. It just means that (if the conditions of the lemmas are fulfilled) no matter which strategy $\mathcal{A}$ we pick, we can always find $x_1, x_2 \in [0, x^{\max}]$ such that $z(\mathcal{A}_{0,S}, x_1) < z(\mathcal{A}, x_1)$ and $z(\mathcal{A}_{0,D}, x_2) < z(\mathcal{A}, x_2)$.

We start with the result that the strategy $\mathcal{A}_{0,S}$ is not weakly dominated if $y > 0$:

Theorem 8.4. *Let* I *be an instance of the TRCP with* $y > 0$*.*

- *For all* $r \in \mathbb{N}$, $\mathcal{A}_{r,S}$ does not weakly dominate $\mathcal{A}_{0,S}$.
- *For all* $r \in \mathbb{N}_0$, $\mathcal{A}_{r,D}$ does not weakly dominate $\mathcal{A}_{0,S}$.

In the next theorem we show that also the strategy $\mathcal{A}_{0,D}$ is not weakly dominated, unless the disruption is over soon (i.e., unless $x^{\max}$ is small).

Theorem 8.5. *Let* I *be an instance of the TRCP.*

- *If* $\max\{\Delta_D + M_D - M_S, \Delta_D\} + y < x^{\max}$, *then for all* $r \in \mathbb{N}_0$, $\mathcal{A}_{r,S}$ does not weakly dominate $\mathcal{A}_{0,D}$.
- *If* $\max\{\Delta_D + M_D - M_S, \Delta_D\} < x^{\max}$, *then for all* $r \in \mathbb{N}$, $\mathcal{A}_{r,D}$ does not weakly dominate $\mathcal{A}_{0,D}$.

8.7.2 Dominance among D-train strategies

This section starts with a characterization of the dominance of a D-train strategy over a later D-train strategy. Lemma 8.6 shows that, if for a D-train strategy $\mathcal{A}_{\tilde{r},D}$ there is no S-train which departs later and arrives earlier than the corresponding D-train, then the D-train strategy $\mathcal{A}_{\tilde{r},D}$ dominates all subsequent D-train strategies $\mathcal{A}_{r,D}$ with $r > \tilde{r}$.

This can be understood as follows: If there is such an S-train, then waiting for the next D-train may have the advantage that possibly the disruption vanishes before the departure of this S-train. In that case, one may take the S-train and arrive earlier than with the first D-train. However, if such an S-train does not exist, then there is no advantage in waiting for the next D-train. The following lemma provides the details.

Lemma 8.6. *Let* I *be an instance of the TRCP, and consider* $\tilde{r} \in \mathbb{N}_0$*. If there does not exist* $r' \in \mathbb{N}_0$ *such that*

$$\mathrm{dep}_D(\tilde{r}) < \mathrm{dep}_S(r') \text{ and } \mathrm{dep}_S(r') + M_S < \mathrm{dep}_D(\tilde{r}) + M_D,$$

then the D-train strategy $\mathcal{A}_{\tilde{r},D}$ weakly dominates *all D-train strategies* $\mathcal{A}_{r,D}$ *with* $r > \tilde{r}$*. Furthermore, the D-train strategy* $\mathcal{A}_{\tilde{r},D}$ dominates *all D-train strategies* $\mathcal{A}_{r,D}$ *with* $r > \tilde{r}$ *and* $\mathrm{dep}_D(r) < x^{\max}$*.*

In Theorem 8.7 we show that the condition of Lemma 8.6 is also *necessary* for dominance among two D-train strategies.

Theorem 8.7. *Let* I *be an instance of the TRCP, and consider* $\tilde{r} \in \mathbb{N}_0$*. Then the D-train strategy* $\mathcal{A}_{\tilde{r},D}$ dominates *all D-train strategies* $\mathcal{A}_{r,D}$ *with* $r > \tilde{r}$ *and* $\mathrm{dep}_D(r) < x^{\max}$ *if and only if there does not exist* $r' \in \mathbb{N}_0$ *such that*

$$\mathrm{dep}_D(\tilde{r}) < \mathrm{dep}_S(r') \text{ and } \mathrm{dep}_S(r') + M_S < \mathrm{dep}_D(\tilde{r}) + M_D. \tag{8.6}$$

8.7.3 Dominance among S-train strategies for the case $M_S < M_D$

In the TRCP, the S-train denotes the train which would arrive earliest at the destination if there was no disruption, i.e., $\Delta_S + M_S < \Delta_D + M_D$. This does not necessarily imply that $M_S < M_D$ However, in this section we explicitly make the assumption that $M_S < M_D$ for discussing dominance among the S-train strategies. The case $M_D \leqslant M_S$ will be treated in Section 8.7.4.

If $M_S < M_D$, then the S-train strategy $\mathcal{A}_{r-1,S}$ weakly dominates the next S-train strategy $\mathcal{A}_{r,S}$ if there does not exist a D-train that departs later than $\mathrm{dep}_S(r-1) + y$ and arrives earlier than $\mathrm{dep}_S(r) + M_S$. Indeed, if such a D-train exists, then, at the departure time $\mathrm{dep}_S(r-1)$ of the S-train, there may be an advantage in waiting for a later train: If the disruption is over after $\mathrm{dep}_S(r-1) + y$ but before the departure of the D-train, then the passenger will arrive earlier by not taking the S-train at time $\mathrm{dep}_S(r-1)$ but a D-train. However, if there is no such D-train, then this advantage does not exist.

Lemma 8.8. *Let* I *be an instance of the TRCP with* $M_S < M_D$*, and consider* $r \in \mathbb{N}_0$*. If there does not exist* $r' \in \mathbb{N}_0$ *such that*

$$\mathrm{dep}_S(r-1) + y < \mathrm{dep}_D(r') \text{ and } \mathrm{dep}_D(r') + M_D < \mathrm{dep}_S(r) + M_S, \tag{8.7}$$

then the S-train strategy $\mathcal{A}_{r-1,S}$ weakly dominates *the S-train strategy* $\mathcal{A}_{r,S}$*. If additionally* $y > 0$ *and* $\text{dep}_S(r-1) < x^{\max}$*, then the S-train strategy* $\mathcal{A}_{r-1,S}$ dominates *the strategy* $\mathcal{A}_{r,S}$.

Note that condition (8.7) of Lemma 8.8 can also be written as follows:

$$\text{dep}_S(r-1) + y < \text{dep}_D(r') < \text{dep}_S(r) + M_S - M_D \tag{8.8}$$

This alternative formulation shows that, if $y \geqslant T_S + M_S - M_D$, then the condition of Lemma 8.8 is always satisfied, independently of any other value. Next we continue with Lemma 8.9, showing that only small additions are needed to make the condition of Lemma 8.8 necessary for dominance among two consecutive S-train strategies.

Lemma 8.9. *Let* I *be an instance of the TRCP with* $M_S < M_D$ *and* $y > 0$*. Furthermore, consider* $r \in \mathbb{N}_0$ *with* $\text{dep}_S(r-1) + y < x^{\max}$*. Then the strategy* $\mathcal{A}_{r-1,S}$ dominates *the strategy* $\mathcal{A}_{r,S}$ *if and only if there does not exist* $r' \in \mathbb{N}_0$ *such that*

$$\text{dep}_S(r-1) + y < \text{dep}_D(r') \textit{ and } \text{dep}_D(r') + M_D < \text{dep}_S(r) + M_S.$$

Now we come to the main result of this section stating that all strategies $\mathcal{A}_{r,S}$ with $r > 0$ are weakly dominated, either by an earlier S-train strategy, or by an earlier D-train strategy. Moreover, if $y > 0$, then all strategies $\mathcal{A}_{r,S}$ with $r > 0$ are dominated.

Theorem 8.10. *If* $M_S < M_D$*, then the S-train strategy* $\mathcal{A}_{r,S}$ *is* weakly dominated *for each* $r \in \mathbb{N}$*. If* $y > 0$*, then the S-train strategy* $\mathcal{A}_{r,S}$ *is* dominated *for each* $r \in \mathbb{N}$ *with* $\text{dep}_S(r) < x^{\max}$.

We will see later in Theorem 8.14 that, if $M_D \leqslant M_S$, then an even stronger statement holds.

8.7.4 The case $M_D \leqslant M_S$

In this section we consider instances of the TRCP with $M_D \leqslant M_S$. However, the assumption $\Delta_S + M_S < \Delta_D + M_D$ from (8.5) still needs to be satisfied. That means that the S-trains do not have a shorter travel time than the D-trains, but, according to the timetable, the first S-train (if it is not hindered) arrives earlier at the destination than the first D-train.

In this section we show that, if $M_D \leqslant M_S$, then the only non-dominated D-train strategy is $\mathcal{A}_{0,D}$. If also $y > 0$, then the only non-dominated S-train strategy is $\mathcal{A}_{0,S}$.

Indeed, Theorem 8.11 shows that in this case $\mathcal{A}_{0,D}$ is the best D-train strategy, since it (weakly) dominates all other D-train strategies. This can be understood as follows: if a D-train is at the point of departing, then waiting for a later D-train can only have an advantage if there is an S-train that departs later than the departing D-train and arrives earlier than this D-train. However, if $M_D \leqslant M_S$, then such an S-train obviously cannot exist.

Theorem 8.11. *Let* I *be an instance of the TRCP with* $M_D \leqslant M_S$. *Then the D-train strategy* $\mathcal{A}_{0,D}$ weakly dominates *the D-train strategy* $\mathcal{A}_{r,D}$ *for all* $r \in \mathbb{N}$. *Moreover,* $\mathcal{A}_{0,D}$ dominates $\mathcal{A}_{r,D}$ *for all* $r \in \mathbb{N}$ *with* $\mathrm{dep}_D(r) < x^{\max}$.

In the remainder of this section we show that, if $M_D \leqslant M_S$ and $y > 0$, then also all S-train strategies $\mathcal{A}_{r,S}$ with $r > 0$ and $\mathrm{dep}_S(r) < x^{\max}$ are dominated. These strategies are dominated either by the S-train strategy $\mathcal{A}_{0,S}$ or by the D-train strategy $\mathcal{A}_{0,D}$. Note that the case $y = 0$ will be treated in Section 8.7.5.

We start with excluding all S-train strategies $\mathcal{A}_{r,S}$ with a small but positive value of r.

Lemma 8.12. *Let* I *be an instance of the TRCP with* $M_D \leqslant M_S$. *Then the S-train strategy* $\mathcal{A}_{0,S}$ weakly dominates *the S-train strategy* $\mathcal{A}_{r,S}$ *for each* $r \in \mathbb{N}$ *such that* r *bounded from above by the relation* $r \leqslant \frac{(\Delta_D+M_D)-(\Delta_S+M_S)}{T_S}$. *If additionally* $y > 0$, *then the S-train strategy* $\mathcal{A}_{0,S}$ dominates *the S-train strategy* $\mathcal{A}_{r,S}$ *for the indicated values of* r.

Next we exclude all S-train strategies $\mathcal{A}_{r,S}$ with a large value of r.

Lemma 8.13. *Let* I *be an instance of the TRCP with* $M_D \leqslant M_S$. *Then the D-train strategy* $\mathcal{A}_{0,D}$ weakly dominates *the S-train strategy* $\mathcal{A}_{r,S}$ *for each* $r \in \mathbb{N}_0$ *with* $r > \frac{(\Delta_D+M_D)-(\Delta_S+M_S)}{T_S}$. *If additionally* $\mathrm{dep}_S(r) < x^{\max}$, *then the D-train strategy* $\mathcal{A}_{0,D}$ dominates *the S-train strategy* $\mathcal{A}_{r,S}$ *for the indicated values of* r.

Combining the results of Theorem 8.11 with the results of Lemmas 8.12 and 8.13, we obtain the main result concerning the case $M_D \leqslant M_S$.

Theorem 8.14. *Let* I *be an instance of the TRCP with* $M_D \leqslant M_S$. *Then all D-train strategies* $\mathcal{A}_{r,D}$ *with* $r \in \mathbb{N}$ *and all S-train strategies* $\mathcal{A}_{r,S}$ *with* $r \in \mathbb{N}$ *are* weakly dominated.

Additionally, all D-train strategies $\mathcal{A}_{r,D}$ *with* $r \in \mathbb{N}$ *and* $\mathrm{dep}_D(r) < x^{\max}$ *are* dominated, *and, if* $y > 0$, *then all S-train strategies* $\mathcal{A}_{r,S}$ *with* $r \in \mathbb{N}$ *and* $\mathrm{dep}_S(r) < x^{\max}$ *are* dominated.

8.7.5 The special case $y = 0$

We now consider the special case that the S-train does not leave the origin station as long as the disruption is present, i.e., $y = 0$. In this case, it is intuitively clear that the passenger cannot gain anything by committing to take an S-train: it is always as good or even better to wait until the disruption is over, and then choose the best option available at that time.

Thus in this case, an S-train strategy belonging to a later S-train (weakly) dominates an S-train strategy belonging to an earlier S-train. Note that this is in contrast with the situation in Lemma 8.9. The foregoing is formalized in the following Lemma.

Lemma 8.15. *Let* I *be an instance of the TRCP with* $y = 0$*, and consider* $r \in \mathbb{N}_0$*. Then the S-train strategy* $\mathcal{A}_{r+1,S}$ weakly dominates *the S-train strategy* $\mathcal{A}_{r,S}$*. If additionally* $\mathrm{dep}_S(r) < x^{\max}$ *and there exists* $r' \in \mathbb{N}_0$ *such that*

$$\mathrm{dep}_S(r) < \mathrm{dep}_D(r') \quad \textit{and} \quad \mathrm{dep}_D(r') + M_D < \mathrm{dep}_S(r+1) + M_S, \tag{8.9}$$

then the S-train strategy $\mathcal{A}_{r+1,S}$ dominates *the S-train strategy* $\mathcal{A}_{r,S}$*.*

In particular, $\mathcal{A}_{\infty,S} = \mathcal{A}_{\infty,D}$ *weakly dominates all S-train strategies and dominates all S-train strategies if* $x^{\max} = \infty$*.*

Finally we describe under which conditions a D-train strategy dominates an S-train strategy if $y = 0$. As in earlier cases, if there exists an S-train that departs later than the involved D-train and arrives earlier than that train, then, at the departure time of the D-train, there may be an advantage in waiting for this S-train. However, if such an S-train does not exist, then the D-train strategy weakly dominates the next S-train strategy. And then it obviously also weakly dominates all later S-train strategies. Furthermore, according to Lemma 8.15, it also dominates all earlier S-train strategies then. The details are provided in Lemma 8.16.

Lemma 8.16. *Let* I *be an instance of the TRCP with* $y = 0$*, and consider* $\tilde{r} \in \mathbb{N}_0$*. If there does not exist* $r' \in \mathbb{N}_0$ *such that*

$$\mathrm{dep}_D(\tilde{r}) < \mathrm{dep}_S(r') \quad \textit{and} \quad \mathrm{dep}_S(r') + M_S < \mathrm{dep}_D(\tilde{r}) + M_D, \tag{8.10}$$

then the D-train strategy $\mathcal{A}_{\tilde{r},D}$ weakly dominates *the S-train strategy* $\mathcal{A}_{r,S}$ *for all* $r \in \mathbb{N}_0$*. If additionally* $x^{\max} > \max\{\mathrm{dep}_D(\tilde{r}), \mathrm{dep}_D(\tilde{r}) + M_D - M_S, \mathrm{dep}_S(r)\}$*, then the D-train strategy* $\mathcal{A}_{\tilde{r},D}$ dominates *the S-train strategy* $\mathcal{A}_{r,S}$ *if and only if condition* (8.10) *is satisfied.*

If $M_D \leqslant M_S$ and $y = 0$, then S-train strategies are also dominated for finite values of $x^{\max}$.

Corollary 8.17. *Let* I *be an instance of the TRCP with* $y = 0$ *and* $M_D \leqslant M_S$*. Furthermore, consider* $\tilde{r}, r \in \mathbb{N}_0$ *such that* $\max\{\mathrm{dep}_D(\tilde{r}), \mathrm{dep}_S(r)\} < x^{\max}$*. Then* $\mathcal{A}_{\tilde{r},D}$ dominates $\mathcal{A}_{r,S}$*.*

8.7.6 Summary

In the previous section we have shown that there exist various dominance relations between strategies. From an algorithmic perspective this is a very useful insight, since we know that we do not need to consider dominated strategies. Hence, as soon as we have decided on a way to measure the quality of the strategies (compare Section 8.5.1), we can find an optimal strategy by enumerating the remaining strategies and computing their objective value.

In Table 8.2, the results of the previous sections are summarized (under the assumption $\Delta_S + M_S < \Delta_D + M_D < \mathrm{suc}_S(x^{\max} - y) + M_S$ which was also made throughout

Results for $M_S < M_D$		Results for $M_D \leqslant M_S$	
Results for $y > 0$			
$\mathcal{A}_{0,D}$ is not weakly dominated	(Theorem 8.5)	$\mathcal{A}_{0,D}$ is not weakly dominated	(Theorem 8.5)
$\mathcal{A}_{0,S}$ is not weakly dominated	(Theorem 8.4)	$\mathcal{A}_{0,S}$ is not weakly dominated	(Theorem 8.4)
if Condition (8.6) holds			
$\mathcal{A}_{r,D}$ is dominated for $r > \tilde{r}$	(Theorem 8.7)	$\mathcal{A}_{r,D}$ is dominated for all $r \in \mathbb{N}$	(Theorem 8.11)
$\mathcal{A}_{r,S}$ is dominated for all $r \in \mathbb{N}$	(Theorem 8.10)	$\mathcal{A}_{r,S}$ is dominated for all $r \in \mathbb{N}$	(Theorem 8.10)
REM=$\{\mathcal{A}_{0,S}, \mathcal{A}_{r,D} \mid 0 \leqslant r \leqslant \tilde{r}\}$	(Equation (8.6))	REM=$\{\mathcal{A}_{0,S}, \mathcal{A}_{0,D}\}$	
Results for $y = 0$			
$\mathcal{A}_{0,D}$ is not weakly dominated	(Theorem8.5)	$\mathcal{A}_{0,D}$ is not weakly dominated	(Theorem8.5)
$\mathcal{A}_{0,S}$ is weakly dominated			
and also dominated			
if $x^{max} = \infty$	(Lemma 8.15)	$\mathcal{A}_{0,S}$ is dominated	(Lemma 8.17)
or if either Condition (8.9) or (8.10) holds	(Lemma's 8.15 and 8.16)		
$\mathcal{A}_{r,D}$ is dominated for $r > \tilde{r}$ if (8.6)	(Theorem 8.7)	$\mathcal{A}_{r,D}$ is dominated for all $r \in \mathbb{N}$	(Theorem 8.11)
$\mathcal{A}_{r,S}$ is weakly dominated for all $r \in \mathbb{N}$			
and also dominated			
if $x^{max} = \infty$	(Lemma 8.15)	$\mathcal{A}_{r,S}$ is dominated for all $r \in \mathbb{N}$	(Lemma 8.17)
or if either Condition (8.9) or (8.10) holds	(Lemma's 8.15 and 8.16)		
REM=$\{\mathcal{A}_{r,D} \mid 0 \leqslant r \leqslant \tilde{r}\}$	(Equation (8.6))	REM=$\{\mathcal{A}_{0,D}\}$	

Table 8.2: Summary of the results from Section 8.7. Only the strategies in the set REM have to be further investigated to find an optimal solution.

Section 8.7). The sets of remaining strategies are denoted by REM in this table, which is short for "Remaining". Note that some entries of the table refer to the value $\tilde{r}$ specified in condition (8.6) in Theorem 8.7. If this condition does not hold for any $\tilde{r}$ we set $\tilde{r} := \infty$.

Note that all strategies with departure times later than x^{max} are equivalent and hence weakly dominate each other. Thus for results of the type "$\mathcal{A}$ is not weakly dominated" or "$\mathcal{A}$ is dominated" we need to assume that x^{max} is large enough, so that the region in which dominance occurs is included in our considerations.

We want to emphasize again that an entry *"not weakly dominated"* means *"not weakly dominated"* and also *"not dominated"*. Furthermore, an entry *"weakly dominated"* means *"weakly dominated"* but not necessarily *"dominated"*.

We see that the result "$\mathcal{A}_{0,D}$ is not weakly dominated" holds in all cases. This can be explained by the fact that, in case of a long disruption, the strategy of taking the first D-train is always best.

Furthermore, we observe that whether y is equal to 0 or larger does not have an influence on the results for the D-train strategies. We see that, if $M_D \leqslant M_S$, then $\mathcal{A}_{r,D}$ is always dominated for $r \in \mathbb{N}$. The reason is that if the disruption has not vanished at time Δ_D, the best strategy is to depart immediately with the D-train, since the travel time with the D-train is shorter than the travel time of the S-train anyway. Note that in fact condition (8.6) in the column "$M_S < M_D$" expresses something similar: Whenever the r-th D-train will arrive earlier than any S-train departing later, waiting does not make sense, hence all D-train strategies with a higher departure time are dominated.

With respect to $\mathcal{A}_{0,S}$, we see that the choice of y matters. If $y > 0$, we see that $\mathcal{A}_{0,S}$ is not weakly dominated. The reason for this result is that the traveler in this case can board the first S-train and travel in it for y minutes. If the disruption vanishes within these y minutes, then he will not have a delay (with respect to his initially planned connection) at all - a result which cannot be achieved with any other strategy. If $y = 0$ on the other hand, then boarding the first S-train at time Δ_S cannot be beneficial, because the train waits at the station until time $\mathrm{suc}_S(x^{max})$. No time advantage can be gained by boarding it, compared to waiting at the station (outside the train) and eventually boarding the train when the disruption is over. On the contrary, we assume that after boarding a train, a passenger cannot get out again. Hence, in this case, $\mathcal{A}_{0,S}$ has a clear disadvantage compared to all strategies which assume the passenger to wait longer. This leads to $\mathcal{A}_{0,S}$ being (weakly) dominated.

Last, we observe that the strategy of taking a later S-train is always weakly dominated and almost always dominated. For the case $y = 0$ it follows from the fact that we cannot gain anything by boarding an S-train before the disruption is over, hence $\mathcal{A}_{r+1,S}$ always dominates $\mathcal{A}_{r,S}$ as long as $x^{max} > \mathrm{dep}_S(r)$ in which case we have weak dominance. For the case $y > 0$ this result is less intuitive. Dominance is ensured by

Theorem 8.10 where we show that for each r, depending on the instance, either an S-train strategy or a D-train strategy which dominates $\mathcal{A}_{r,S}$ can be found.

Altogether, we find for the special case $M_D \leqslant M_S$ and $y = 0$ that $\mathcal{A}_{0,D}$ is the only remaining strategy - it will be optimal for any reasonable way of measuring solution quality under uncertainty.

If $M_D \leqslant M_S$ and $y > 0$, only $\mathcal{A}_{0,D}$ and $\mathcal{A}_{0,S}$ remain. Any reasonably acting traveler should choose among these two strategies. To find the best strategy for a fixed quality measure under uncertainty, we can evaluate both and compare the results.

For the case $M_S < M_D$ and $y = 0$, then all remaining strategies are D-train strategies. That means that the traveler should not board an S-train unless the disruption has vanished. Which of the D-train strategies is best for this uncertain solution will depend on the definition of measuring the solution quality. If there is a number $\tilde{r}$ as given by (8.6), the number of strategies is finite and can be evaluated and compared. But even if the set REM as defined in the table is infinite because there does not exist such an $\tilde{r}$, it is to expect that either $\mathcal{A}_{\infty,D}$ is optimal or $\mathcal{A}_{r,D}$ is increasing in r from an early point on.

The case $M_S < M_D$ and $y > 0$ covers the example from Section 8.5.2. Here, we have the highest dependence of "optimality" on the chosen measure to evaluate objective quality under uncertainty. With respect to D-train strategies, this case is similar to the previous ones. However, the traveler should as well consider taking the first S-train, since this strategy is not weakly dominated either.

8.8 Discussion and Conclusion

In this chapter we discussed the Traveler's Route Choice Problem (TRCP). This is the problem a traveler faces in a periodic railway system if the fastest route to his destination is blocked by a disruption. He then has the option to wait until the disruption is over, or to take a train along a detour route. The TRCP is a typical example of an optimization problem under uncertainty, where some of the parameters of the problem, in this case the duration of the disruption, are not known in advance.

Since in an instance of the TRCP the duration of the disruption is unknown, it is impossible to say in advance which routing option will lead to the earliest arrival time. We hence propose the concepts of *dominance* and *weak dominance* among the practical strategies: A strategy is said to (weakly) dominate another strategy if using the first strategy will not lead to a later arrival than using the second one, independently of the actual duration of the disruption. This implies that the first strategy is never worse than the second strategy, no matter which of the quality measures described in Section 8.5.1 is used. We hence may ignore dominated strategies, which reduces computation time when solving the robust, stochastic, or online optimization problem.

Applied to the TRCP, we have shown that the strategy to take the first D-train is always non-dominated. Furthermore, taking an S-train can only be beneficial if the S-train can drive on until it reaches the disruption. When the S-train has to wait at the origin station, S-train strategies are dominated.

While we are able to identify non-dominated strategies, we have not answered the relevant question which of the non-dominated strategies leads to the earliest arrival time in the worst case, for the case that regret or competitive ratio is considered, or on average. We decided to address this question in future research.

We want to conclude this chapter with noting that the concept of dominance that we described here in the context of the TRCP can be applied directly to other optimization problems under uncertainty as well. However, the TRCP is an excellent problem to illustrate these concepts, since in this problem the uncertainty that is to be handled is restricted to just a single parameter, namely the end time of the disruption.

References

W. Adamowicz, J. Louviere, and M. Williams. Combining revealed and stated preference methods for valuing environmental amenities. *Journal of Environmental Economics and Management*, 26(3):271–292, 1994.

B. Agard, C. Morency, and M. Trépanier. Mining public transport user behaviour from smart card data. In *12th IFAC Symposium on Information Control Problems in Manufacturing-INCOM*, pages 17–19. 2006.

W. Allen. Stardust memories. 35mm Black and White film, 1980. Directed by Woody Allen. Cinematrography by Gordon Willis. Production company: Rollins-Joffe Productions.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

W. B. Arthur. Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, 1994.

E. Avineri and J. N. Prashker. Sensitivity to uncertainty: need for a paradigm shift. *Transportation Research Record: Journal of the Transportation Research Board*, 1854(1):90–98, 2003.

K. Axhausen, M. Balmer, and F. Ciari. A new mode choice model for a multi-agent transport simulation. In *8th Swiss Transport Research Conference. ETH, Eidgenössische Technische Hochschule*. 2008.

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

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

M. Balmer, K. Axhausen, and K. Nagel. Agent-based demand-modeling framework for large-scale microsimulations. *Transportation Research Record: Journal of the Transportation Research Board*, 1985:125–134, 2006.

A. Bar-Noy and B. Schieber. The Canadian Traveller Problem. In *Proceedings of the 2nd Annual ACM-SIAM Symposium on on Discrete Algorithms*. 1991.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the Third International Conference on Weblogs and Social Media*. 2009. URL `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

A. L. Bazzan, R. H. Bordini, G. K. Andrioti, R. M. Vicari, and J. Wahle. Wayward agents in a commuting scenario (personalities in the minority game). In *Proceedings of the Fourth International Conference on MultiAgent Systems*, pages 55–62. IEEE, 2000.

D. E. Bell. Regret in decision making under uncertainty. *Operations Research*, 30(5):961–981, 1982.

M. Ben-Akiva, M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. Combining revealed and stated preferences data. *Marketing Letters*, 5(4):335–349, 1994.

M. Ben-Akiva and S. Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. The MIT press, 1985.

N. Ben-Khedher, J. Kintanar, C. Queille, and W. Stripling. Schedule optimization at SNCF: From conception to day of departure. *Interfaces*, pages 6–23, 1998.

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

M. Bierlaire. Discrete choice models. In *Operations Research and Decision Aid Methodologies in Traffic and Transportation Management*, pages 203–227. Springer, 1998.

J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer Verlag, New York, 1997. ISBN 0387982175.

N. Boccara. *Modeling complex systems*. Springer Science & Business Media, 2010.

R. Borndörfer, M. Grötschel, and M. E. Pfetsch. A column-generation approach to line planning in public transport. *Transportation Science*, 41(1):123–132, 2007.

A. Borodin and R. El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 1998.

P. Bouman. Synthetic Smart Card Data for the Analysis of Temporal and Spatial Patterns. 2017. URL `http://dx.doi.org/10.5281/zenodo.321686`.

J. Bowman and M. Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1):1–28, 2001.

C. Büsing. The exact subgraph recoverable robust shortest path problem. In R. Ahuja, R. Möhring, and C. Zaroliagis, editors, *Robust and Online Large-Scale Optimization*, volume 5868 of *Lecture Notes in Computer Science*, pages 231–248. Springer, Heidelberg, 2009. ISBN 978-3-642-05464-8.

C. Büsing. Recoverable robust shortest path problems. *Networks*, 59:181–189, 2012.

G. D. B. Cameron and G. I. D. Duncan. Paramics - parallel microscopic simulation of road traffic. *The Journal of Supercomputing*, 10:25–53, 1996. ISSN 0920-8542. URL `http://dx.doi.org/10.1007/BF00128098`. 10.1007/BF00128098.

R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):245–276, 1966.

D. Challet, M. Marsili, and R. Zecchina. Statistical mechanics of systems with heterogeneous agents: Minority games. *Physical Review Letters*, 84(8):1824, 2000.

D. Challet, M. Marsili, and Y.-C. Zhang. Minority games: interacting agents in financial markets. *OUP Catalogue*, 2004.

J.-P. Chancelier, M. De Lara, and A. de Palma. Risk aversion in expected intertemporal discounted utilities bandit problems. *Theory and Decision*, 67(4):433–440, 2009.

D. Charypar and K. Nagel. Generating complete all-day activity plans with genetic algorithms. *Transportation*, 32:369–397, 2005. ISSN 0049-4488. URL `http://dx.doi.org/10.1007/s11116-004-8287-y`. 10.1007/s11116-004-8287-y.

C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68:285–299, 2016.

M. S. Daskin, S. M. Hesse, and C. S. Revelle. α-reliable p-minimax regret: a new model for strategic facility location modeling. *Location Science*, 5(4):227–246, 1997.

A. De Palma, R. Lindsey, and N. Picard. Risk aversion, the value of information, and traffic equilibrium. *Transportation Science*, 46(1):1–26, 2012.

F. Devillaine, M. Munizaga, and M. Trépanier. Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, 2276(1):48–55, 2012.

L. C. Dias and J. N. Climaco. Shortest path problem with partial information: models and algorithms for detecting dominance. *European Journal of Operational Research*, 121:16–31, 2000.

J. Dibbelt, T. Pajor, B. Strasser, and D. Wagner. Intriguingly simple and fast transit routing. In *International Symposium on Experimental Algorithms*, pages 43–54. Springer, 2013.

A. Downey. *Think Complexity: Complexity Science and Computational Modeling*. O'Reilly & Associates Inc, 2012.

N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

M. Ehrgott. *Multicriteria optimization*, volume 2. Springer, 2005.

F. Fellini. Otto e Mezzo. 35mm Black and White film, 1963. Directed by Federico Fellini. Cinematography by Gianni Di Venanzo. Production companies: Cineriz; Francinex.

P. Fioole, L. Kroon, G. Maróti, and A. Schrijver. A rolling stock circulation model for combining and splitting of passenger trains. *European Journal of Operational Research*, 174(2):1281–1297, 2006.

P. C. Fishburn. The foundations of expected utility. *Theory & Decision Library*, 1982.

G. Flötteröd. A search acceleration method for optimization problems with transport simulation constraints. *Transportation Research Part B: Methodological*, 98:239 – 260, 2017. ISSN 0191-2615.

G. Flötteröd, M. Bierlaire, and K. Nagel. Bayesian demand calibration for dynamic traffic simulations. *Transportation Science*, 45(4):541–561, 2011.

L. R. Ford and D. R. Fulkerson. *Flows in networks*. Princeton university press, 1962.

H. Fujiwara and K. Iwama. Average-case competitive analyses for ski-rental problems. In *Algorithms and Computation*, pages 476–488. Springer, 2002.

V. Gabrel and C. Murat. Robust shortest path problems. Technical report, LAMSADE, Université Paris-Dauphine, 2007. URL `http://hal.archives-ouvertes.fr/hal-00179975/`.

A. Galstyan, S. Kolar, and K. Lerman. Resource allocation games with changing resource capacities. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 145–152. ACM, 2003.

Y. Gao. Shortest path problem with uncertain arc lengths. *Computers and Mathematics with Applications*, 62:2591–2600, 2011.

M. Gardner. Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223:120–123, 1970.

S. Gast. *Network Flow Problems with Uncertain Input Data in the Context of Supply Chain Management Applications*. Ph.D. thesis, University of Wuppertal, 2010.

M. Goerigk, M. Knoth, M. Müller-Hannemann, M. Schmidt, and A. Schöbel. The price of robustness in timetable information. *Transportation Science*, 2013a. Accepted for publication.

M. Goerigk, M. Schmidt, A. Schöbel, M. Knoth, and M. Müller-Hannemann. The price of strict and light robustness in timetable information. *Transportation Science*, 48(2):225–242, 2013b.

D. Golovin, V. Goyal, V. Polishchuk, R. Ravi, and M. Sysikaski. Improved approximations for two-stage min-cut and shortest path problems under uncertainty. *Mathematical Programming*, pages 1–28, 2014.

R. M. Goodwin. A growth cycle. In *Essays in Economic Dynamics*, pages 165–170. Springer, 1982.

R. Hites, Y. De Smet, N. Risse, M. Salazar-Neumann, and P. Vincke. About the applicability of MCDA to some robustness problems. *European Journal of Operational Research*, 174(1):322–332, 2006.

A. Horni, K. Nagel, and K. Axhausen, editors. *The Multi-Agent Transport Simulation MATSim*. London: Ubiquity Press, 2016. URL `http://dx.doi.org/http://dx.doi.org/10.5334/baw`.

B. A. Huberman and R. M. Lukose. Social dilemmas and internet congestion. *Science*, 277(5325):535–537, 1997.

E. van der Hurk. *Passengers, Information, and Disruptions*. Ph.D. thesis, Erasmus Research Institute of Management (ERIM), 2015.

E. van der Hurk, L. Kroon, and G. Maróti. Passenger advice and rolling stock rescheduling under uncertainty for disruption management. *Transportation Science, forthcoming*, 2017.

E. van der Hurk, L. Kroon, G. Maróti, and P. Vervest. Deduction of passengers' route choices from smart card data. *Intelligent Transportation Systems, IEEE Transactions on*, 16(1):430–440, 2015.

D. A. Iancu and N. Trichakis. Pareto efficiency in robust optimization. *Management Science*, 60(1):130–147, 2014.

I. Kaddoura, B. Kickhöfer, A. Neumann, and A. Tirachini. Agent-based optimisation of public transport supply and pricing: Impacts of activity scheduling decisions and simulation randomness. *Transportation*, 42(6):1039–1061, 2015.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

R. M. Karp. Reducibility Among Combinatorial Problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

W. Kets. Learning with fixed rules: The minority game. *Journal of Economic Surveys*, 26(5):865–878, 2012.

M. Keyhani, M. Schnee, and K. Weyhe. Path finding strategies in stochastic networks. Technical report, TU Darmstadt, 2014.

K. Klamroth, E. Köbis, A. Schöbel, and C. Tammer. A unified approach for different concepts of robustness and Stochastic Programming via nonlinear scalarizing functionals. *Optimization*, 62(5):649–671, 2013.

K. Klamroth, E. Köbis, A. Schöbel, and C. Tammer. A unified approach for uncertain optimization. 2015. Working paper.

F. H. Knight. *Risk, uncertainty and profit*. University of Chicago Press, 1921.

J. Knockaert, Y.-Y. Tseng, E. T. Verhoef, and J. Rouwendal. The "spitsmijden" experiment: A reward to battle congestion. *Transport Policy*, 24:260–272, 2012.

P. Kouvelis and S. Sayin. Algorithm robust for the bicriteria discrete optimization problem. *Annals of Operations Research*, 147:71–85, 2006.

P. Kouvelis and G. Yu. *Robust Discrete Optimization and Its Applications*. Kluwer Academic Publishers, 1997.

L. Kroon and P. Vervest. Complexity in public transport: Passenger behavior and system optimization. 2010. Grant Proposal, number 645.000.008.

T. C. Lam and K. A. Small. The value of time and reliability: measurement from a value pricing experiment. *Transportation Research Part E: Logistics and Transportation Review*, 37(2):231–251, 2001.

H. Levy. Stochastic dominance and expected utility: survey and analysis. *Management Science*, 38(4):555–593, 1992.

Z. Li and D. A. Hensher. Crowding and public transport: A review of willingness to pay evidence and its relevance in project appraisal. *Transport Policy*, 18(6):880 – 887, 2011. ISSN 0967-070X.

H. Link. Pep–a yield-management scheme for rail passenger fares in Germany. *Japan Railway & Transport Review*, 38(3):50–55, 2004.

Y. Liu, J. Bunker, and L. Ferreira. Transit users' route-choice modelling in transit assignment: A review. *Transport Reviews*, 30(6):753–769, 2010.

S. Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

G. Loomes and R. Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, pages 805–824, 1982.

E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.

A. J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3):271–274, 1910.

A. Lubach. Zondag met Lubach, Season 3 Episode 3. Television, 2015. Production companies: Human Factor and VPRO.

M. G. McNally. The four step model. In D. Hensher and K. Button, editors, *Handbook of Transport Modelling*, Handbooks in Transportation Research Series. Elsevier, 2000.

K. Meister, M. Balmer, F. Ciari, A. Horni, M. Rieser, R. Waraich, and K. Axhausen. Large-scale agent-based travel demand optimization applied to Switzerland, including mode choice. In *12th World Conference on Transportation Research, Lisbon, July 2010*. 2010.

E. Miller. Microsimulation and activity-based forecasting. In *Activity-Based Travel Forecasting Conference*. 1997.

C. Morency, M. Trépanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

K. Müller and K. Axhausen. Population synthesis for microsimulation: State of the art. Technical report, ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT), 2010.

K. Nagel and M. Schreckenberg. A cellular automaton model for freeway traffic. *Journal de physique I*, 2(12):2221–2229, 1992.

J. von Neumann and O. Morgenstern. Game theory and economic behavior. *Princeton, Princeton University*, 1944.

Y. M. Nie and X. Wu. Shortest path problem considering on-time arrival probability. *Transportation Research Part B: Methodological*, 43(6):597–613, 2009.

NOS. Tijdelijk minder treinen door opdoeken oude rijtuigen. *Website of the Netherlands Broadcasting Foundation (NOS), NOS Nieuws*, 2015. URL `http://nos.nl/artikel/2059442-tijdelijk-minder-treinen-door-opdoeken-oude-rijtuigen.html`.

NWO. Grip on complexity: How manageable are complex systems? Technical report, Netherlands Organisation for Scientific Research (NWO), 2014.

K. Oosterhuis and W. van Meteren. Studenten in verzet tegen NS. *Trouw*, 2016. URL `http://www.trouw.nl/tr/nl/4324/Nieuws/article/detail/4258055/2016/03/07/Studenten-in-verzet-tegen-NS.dhtml`.

OpenStreetMap Foundation. Openstreetmap. 2012. URL `http://www.openstreetmap.org`. Http://www.openstreetmap.org.

Y. Pan, L. Sun, and M. Ge. Finding reliable shortest path in stochastic time-dependent network. *Procedia-Social and Behavioral Sciences*, 96:451–460, 2013.

C. Papadimitriou and M. Yannakakis. Shortest paths without a map. *Theoretical computer science*, 84(1):127, 1991.

M.-P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557 – 568, 2011. ISSN 0968-090X. URL http://dx.doi.org/10.1016/j.trc.2010.12.003.

P. Perny and O. Spanjaard. An axiomatic approach to robustness in search problems with multiple scenarios. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 469–476. Morgan Kaufmann Publishers Inc., 2002.

M. Rieser. *Adding transit to an agent-based transportation simulation*. Ph.D. thesis, Technical University Berlin, Berlin, 2010.

C. Rommel. Automatic feedback control applied to microscopically simulated traffic the potential of route guidance in the Berlin traffic network. Technical report, VSP Working Paper, 2007.

R. W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

R. Savit, R. Manuca, and R. Riolo. Adaptive competition, market efficiency, and phase transitions. *Physical Review Letters*, 82(10):2203, 1999.

T. C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2):143–186, 1971.

A. Schöbel. Line planning in public transportation: models and methods. *OR Spectrum*, 34(3):491–510, 2012.

J. Schumacher. Falling down. 35mm Technicolor film, 1993. Directed by Joel Schumacher. Cinematography by Andrzej Bartkowiak. Production companies: Alcor Films, Canal+, Regency Enterprises and Warner Bros.

R. Selten, T. Chmura, T. Pitz, S. Kube, and M. Schreckenberg. Commuters route choice behaviour. *Games and Economic Behavior*, 58(2):394–406, 2007.

C. E. Sigal, A. A. B. Pritsker, and J. J. Solberg. The stochastic shortest route problem. *Operations Research*, 28(5):1122–1129, 1980.

S. S. Skiena. *The Algorithm Design Manual*. Springer, 2nd edition, 2008. ISBN 1848000693.

D. D. Sleator and R. E. Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.

K. Talluri and G. Van Ryzin. *The theory and practice of revenue management*, volume 68. Springer Verlag, 2005.

The Apache Software Foundation. Math - Commons-Math: The Apache Commons Mathematics Library. 2012. URL `http://commons.apache.org/math/`. Version 3.0.

M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805, 2000.

M. Trépanier, N. Tranchant, and R. Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14, 2007.

R. Tsuchiya, Y. Sugiyama, K. Yamauchi, K. Fujinami, R. Arisawa, and T. Nakagawa. Route-choice support system for passengers in the face of unexpected disturbance of train operations. In *Computers in Railways X. The Tenth International Conference*. 2006.

M. A. Valdebenito and G. I. Schuëller. A survey on approaches for reliability-based optimization. *Structural and Multidisciplinary Optimization*, 42(5):645–663, 2010.

W. Vickrey. Pricing in urban and suburban transport. *The American Economic Review*, pages 452–465, 1963.

J. Wahle, A. Bazzan, F. Klügl, and M. Schreckenberg. Decision dynamics in a traffic scenario. *Physica A: Statistical Mechanics and its Applications*, 287(3):669–681, 2000.

J. Wahle, A. Bazzan, F. Klügl, and M. Schreckenberg. The impact of real-time information in a two-route scenario using agent-based simulation. *Transportation Research Part C: Emerging Technologies*, 10(5):399–417, 2002.

M. Wardman. A comparison of revealed preference and stated preference models of travel behaviour. *Journal of Transport Economics and Policy*, pages 71–91, 1988.

J. Wardrop. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers*, 1:325–378, 1952.

D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393(6684):440–442, 1998.

S. Westphal. A note on the k-Canadian traveler problem. *Information Processing Letters*, 106(3):87–89, 2008.

G. Yu and J. Yang. On the robust shortest path problem. *Computers and Operations Research*, 25(6):457–468, 1998.

M. Zou, M. Li, X. Lin, C. Xiong, C. Mao, C. Wan, K. Zhang, and J. Yu. An agent-based choice model for travel mode and departure time and its case study in beijing. *Transportation Research Part C: Emerging Technologies*, 2015.

Appendices

A

An introduction to k-means clustering

The common goal of all methods for *cluster analysis* is to partition data (points, observations, records, etc.) into groups of objects that are similar to each other or share similar properties (a *cluster* is just a fancy name for a group of such objects). This way large datasets can be easily summarised, as we only need to study one representative object per group instead of the entire dataset. For example, we can reduce the profiles of a million passengers into a few profiles of prototype passengers.

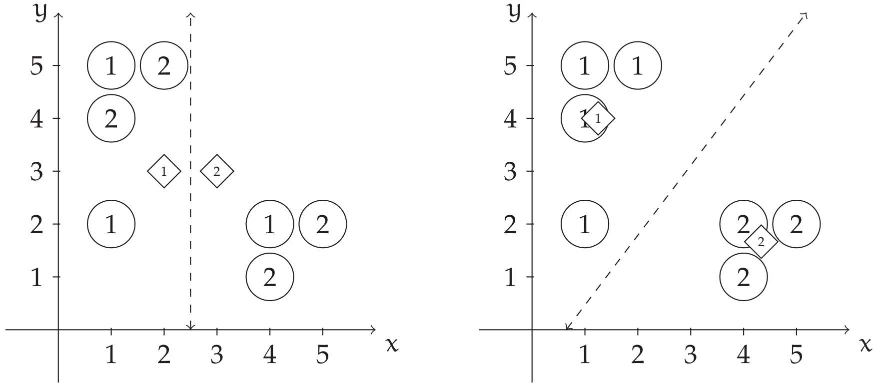
Although there are many different clustering methods available, we use one of the most popular and simple ones in this thesis: k-means clustering. The k refers to a fixed number of clusters that we have to decide before we feed our data to the clustering method, and *means* refers to the principle that the representative of each cluster is the *mean* (or average) of the objects in that cluster. This implies that there exists a way to compute an average for any group of objects. As a result, objects are often interpreted as vectors in an Euclidean space within many practical applications. The objective is to assign objects to clusters such that the distance between objects in the same cluster is small, while the distance between objects of different clusters is large.

As in general it is NP-hard to find a clustering that minimizes the within-cluster distance, we usually resort to a simple, iterative procedure formally known as the algorithm of Lloyd (Lloyd, 1982), but often called the k-means algorithm. The procedure produces reasonably good clusterings in some applications and also runs fast on large datasets. In situations where the algorithm runs into a bad local optimum, it may be helpful to look at different clustering approaches. The k-means clustering algorithm can be described in a few sentences. Initially, each object in the dataset is assigned to a random cluster among the k target clusters. We then iterate between two steps: (1) compute the average, often called a centroid, for each cluster based on the current assignment of objects to clusters, (2) assign each object to the cluster of which centroid computed in step 1 is the closest to the object. If we do not update any cluster assignments in step 2, the algorithm has converged and we can stop. Alternatively we can stop after a predefined number of iterations has been performed. As both steps can be performed in $O(nk)$ time, there is a sound theoretical argument that explains why the procedure runs fast on large datasets. Furthermore, if the initial random assignment of objects to clusters is done carefully (Arthur and Vassilvitskii, 2007) the procedure is guaranteed to converge to a solution that is at most a factor $O(\log k)$ away from the optimal clustering. This procedure is referred to as the k-means++ algorithm and avoids the issues of picking a random initial clustering that results in a poor within-cluster distance. An open-source implementation of this algorithm is available in the Apache Commons Math library (The Apache Software Foundation, 2012).

In Figure A.1 we visualise two iterations of the algorithm applied to a small dataset, in this case for two clusters. In Figure A.1a we see the initial random assignment of clusters to the data objects, indicated by the numbers in the circular points of the dataset. The centroids are indicated by the diamonds and can be computed by taking the average for each dimension of the points in the cluster. Based on the centroid, we can compute the regions that are closest to one of them. In this case, this is a single line separating the plane. For larger values of k, this separation can be visualised using a *voronoi* diagram. In Figure A.1b the algorithm has updated the clusters associated with each node; all nodes that were in the region of centroid 1 are now assigned to cluster 1, and similarly for cluster 2. As a result, the centroids move to new positions and the boundary between the clusters changes. As the updated boundary perfectly separates the objects of cluster 1 and cluster 2, the assignment of objects to clusters will not change and we can conclude the algorithm has converged.

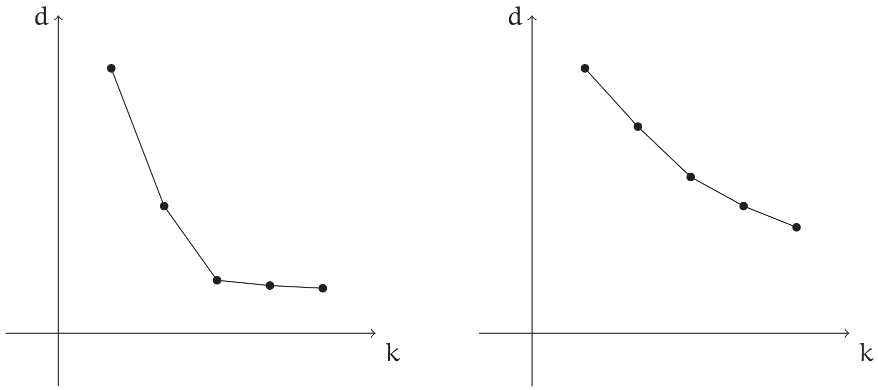
One of the main drawbacks of the k-means clustering method is that it is not always obvious how to select k, as this needs to be defined prior to the application of the method to the dataset. Although there are various ways to overcome this issue, a simple approach is to try a number of different values of k and investigate the quality of the resulting solution. A popular quality measure is the within-cluster distance: The average distance between all pairs of objects in a single cluster. When this measure is small, we know that a cluster covers a (relatively) small area that is densely populated with objects. If we increase the number of available clusters, we can always split up the cluster where the average distance is the greatest into two smaller clusters. Thus, we can see that the optimal within-cluster distance decreases monotonically in the number of clusters. Note that this decrease may not always be observed when we apply the k-means algorithm, as it is not guaranteed to find the optimal clustering, but such situations are not very common in practice.

A quick and popular way to investigate the quality of every k is to draw a scree-plot (Cattell, 1966), sometimes referred to as an elbow plot. Such a plot consists of



(a) First iteration (b) Second and final iteration

Figure A.1: Example of the k-means clustering algorithm. The objects are visualised by circles containing the number of the cluster they are currently assigned to while the cluster centroids are shown as diamonds. The dashed line indicates the boundary between the two clusters based on the current cluster centroids.



(a) Helpful scree plot (b) Less helpful scree plot

Figure A.2: Two example scree plots that plot the number of clusters k against the average within-cluster distance d, which allows us to investigate a good choice for k. Typically we prefer to select the number of clusters in such a way that adding more clusters results only in a marginal decrease of the within-cluster distance. If such a situation occurs, the plot contains a sharp bend. It is not guaranteed that such a situation occurs.

a line plot with different values for k on the horizontal axis and the within-cluster distance of the vertical axis. If the plot shows a sharp bend, the value of k before or associated with the corner of the bend is typically chosen as the preferred value, as further increases in k only result in marginal improvements of the within-cluster distance. While the advantage of this approach is that it is easy to understand, it has the disadvantage that we do not always see a nice elbow-like shape in a scree plot. If that is the case, different criteria such as the interpretability of results, can also be considered before settling for a particular choice of k.

In Figure A.2 we show two example scree plots that one could observe in practice. In the scree plot of Figure A.2a, we see a sharp bend at the third value of k, and thus this is an obvious choice based on the visualised scree-plot. However, something like Figure A.2b can also occur, resulting in no clear winner for the preferred value of k. In both cases, it is advisable to have a critical look at the interpretability of the resulting clusters, but as the scree-plot is not helpful it becomes a lot more important in the second case.

B

Appendices to the Survey Experiment

B.1 Instructions for Respondents

Introduction In this experiment, you will have to choose how you will travel to your work. You can go by train or by car. If you go by train, you can check the train schedule to choose the best train for you. There are many things to consider, but you should take the option which makes the most sense to you.

Structure The experiment consists of three phases and you will be guided through those phases. In the first two phases you will have to make a travel decision. After the trip you can evaluate how satisfied you are with the result. There are 20 rounds per phase, so a total of 40 rounds. At the conclusion of the phase, you will be given a summary of your travel outcomes. The third and final phase is a follow-up survey.

Instruction In this phase, you have to travel from Utrecht to Den Bosch for your work. Your work is situated at the back of the station, so that the travel time from the station to your workplace is negligible. Every morning you have to make a decision on how you are going to travel to work. Every mode of traveling can have its advantages and disadvantages. For travel by train there can be disruptions, for travel by car there can be traffic jams. If there is enough room in the train you can sit comfortably and do some work. If you are able to work in the train, you can go home sooner at the end of the day. Every minute that you leave later than the departure time of the first option allows you to sleep in longer. However, your manager has instructed you to be at your work at 8:30. You decide yourself which of these aspects are the most important to you. This phase has 20 rounds. Every round has two decisions: You first choose the trip you wish to take, and after seeing the result your satisfaction with the particular trip. At the end of this phase, you will be presented with a summary of your travel outcomes, which includes your total travel time, total time of delay, number of times you were able to work in the train, the number of times you had to stand in the train, the number of disruptions, and the number of late arrivals. You are not the only one making these decisions. Your fellow travelers will also make choices based on their travel experiences, which influences how crowded particular trains and the road will be.

Added Instructions for Second Phase This phase is almost the same as the last phase. You can re-read the instructions for that phase below. The only addition is that you can use the mobile app in the morning to see how crowded the train company thinks the train will be. Other travelers also have access to this information so that it can also influence their travel behavior.

B.2 Group Comparisons of Measurements

On the following pages we present additional tables that are useful for the comparision of the different treatment groups. In each table the comparison between two treatment groups is split out between the the two phases of the quasi-experiment and the difference between the two phases. These comparisions are given in Tables B.1, B.2 andB.3. Figure B.1 contains additional visualizations of descriptive statistics observed during the quasi-experiment.

Table B.1: Comparison of the measures when the group of respondents is splitted according to the quality of information q_j

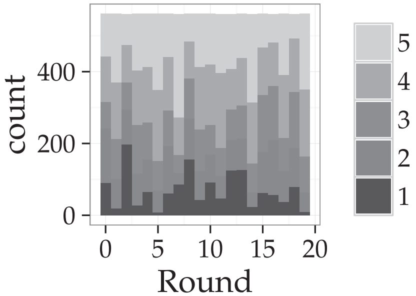
	Information = 0		Information = 1			
	Mean	Std. Dev	Mean	Std. Dev	Diff	p-value
ca-1	0.32	0.19	0.36	0.20	0.04	0.03
ls-1	0.62	0.38	0.68	0.37	0.05	0.11
$mtsc_2$-1	0.38	0.37	0.44	0.38	0.06	0.06
sp-1	0.38	0.31	0.44	0.31	0.06	0.03
rsc-1	0.38	0.31	0.44	0.31	0.06	0.04
rsca-1	0.58	0.16	0.60	0.17	0.02	0.21
avg_S-1	0.56	0.08	0.56	0.08	0.01	0.47
maxstr-1	0.48	0.33	0.42	0.32	-0.06	0.04
dr-1	0.52	0.15	0.52	0.15	0.00	0.94
ca-2	0.19	0.16	0.48	0.19	0.29	0.00
ls-2	0.86	0.23	0.80	0.26	-0.06	0.01
$mtsc_2$-2	0.57	0.36	0.53	0.36	-0.04	0.17
sp-2	0.66	0.27	0.52	0.24	-0.14	0.00
rsc-2	0.66	0.27	0.51	0.24	-0.15	0.00
rsca-2	0.54	0.16	0.63	0.13	0.10	0.00
avg_S-2	0.65	0.09	0.56	0.08	-0.09	0.00
maxstr-2	0.26	0.23	0.30	0.25	0.04	0.03
dr-2	0.43	0.13	0.52	0.13	0.08	0.00
si-2	0.42	0.22	0.38	0.23	-0.04	0.07
Φ_{ca}	-0.13	0.22	0.13	0.23	0.26	0.00
Φ_{ls}	0.23	0.39	0.12	0.33	-0.11	0.00
Φ_{mtsc_2}	0.19	0.39	0.08	0.38	-0.11	0.00
Φ_{sp}	0.28	0.31	0.08	0.25	-0.20	0.00
Φ_{rsc}	0.28	0.31	0.07	0.26	-0.20	0.00
Φ_{rsca}	-0.05	0.24	0.03	0.18	0.08	0.00
Φ_{avg_S}	0.09	0.09	-0.01	0.08	-0.09	0.00
Φ_{maxstr}	-0.22	0.31	-0.12	0.27	0.11	0.00
Φ_{dr}	-0.09	0.17	-0.00	0.15	0.08	0.00

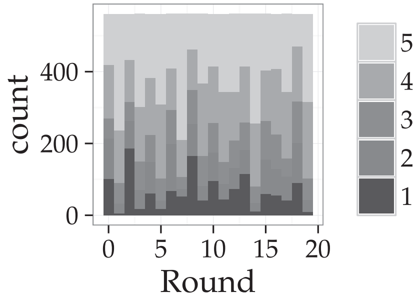
Table B.2: Comparison of the measures when the group of respondents is slitted according to the occurrence of large disruptions δ_j

	Disruption = 0		Disruption = 1			
	Mean	Std. Dev	Mean	Std. Dev	Diff	p-value
ca-1	0.34	0.18	0.34	0.21	0.01	0.69
ls1	0.69	0.35	0.61	0.40	-0.08	0.02
$mtsc_2$-1	0.44	0.37	0.38	0.38	-0.06	0.07
sp-1	0.42	0.29	0.39	0.32	-0.03	0.25
rsc-1	0.43	0.29	0.40	0.33	-0.03	0.23
rsca-1	0.62	0.15	0.57	0.17	-0.05	0.00
avg_S-1	0.54	0.08	0.58	0.09	0.03	0.00
maxstr-1	0.41	0.31	0.49	0.34	0.08	0.01
dr-1	0.51	0.13	0.53	0.17	0.03	0.03
ca-2	0.30	0.24	0.37	0.21	0.07	0.00
ls-2	0.85	0.22	0.81	0.27	-0.03	0.12
$mtsc_2$-2	0.57	0.35	0.53	0.37	-0.04	0.17
sp-2	0.61	0.26	0.58	0.27	-0.03	0.23
rsc-2	0.60	0.26	0.57	0.28	-0.03	0.24
rsca-2	0.59	0.15	0.58	0.17	-0.01	0.54
avg_S-2	0.59	0.09	0.62	0.10	0.03	0.00
maxstr-2	0.26	0.22	0.30	0.26	0.04	0.04
dr-2	0.49	0.12	0.46	0.14	-0.03	0.00
si-2	0.44	0.24	0.37	0.21	-0.07	0.00
Φ_{ca}	-0.04	0.26	0.02	0.26	0.06	0.01
Φ_{ls}	0.16	0.33	0.20	0.39	0.05	0.15
Φ_{mtsc_2}	0.13	0.39	0.15	0.39	0.02	0.65
Φ_{sp}	0.18	0.29	0.18	0.31	0.00	0.92
Φ_{rsc}	0.17	0.29	0.18	0.32	0.00	0.86
Φ_{rsca}	-0.03	0.19	0.01	0.23	0.04	0.04
Φ_{avg_S}	0.04	0.10	0.04	0.10	-0.01	0.51
Φ_{maxstr}	-0.15	0.28	-0.19	0.30	-0.04	0.15
Φ_{dr}	-0.01	0.15	-0.08	0.18	-0.06	0.00

Table B.3: Comparison of the measures when the group of respondents is slitted according to the reactivity of the crowd ρ_j

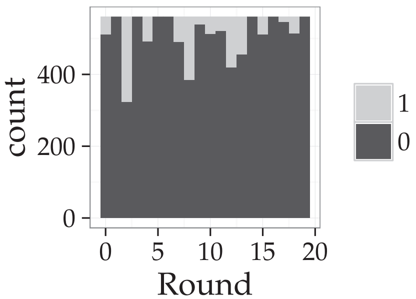
	Reactive cap. = 0		Reactive cap. = 1			
	Mean	Std. Dev	Mean	Std. Dev	Diff	p-value
ca-1	0.37	0.18	0.31	0.20	-0.06	0.00
ls-1	0.59	0.39	0.70	0.35	0.11	0.00
$mtsc_2$-1	0.37	0.37	0.45	0.38	0.08	0.01
sp-1	0.37	0.31	0.44	0.30	0.07	0.01
rsc-1	0.37	0.31	0.45	0.31	0.08	0.01
rsca-1	0.62	0.13	0.57	0.19	-0.05	0.00
avg_S-1	0.58	0.08	0.54	0.08	-0.04	0.00
maxstr-1	0.50	0.34	0.41	0.31	-0.09	0.00
dr-1	0.54	0.15	0.50	0.15	-0.04	0.01
ca-2	0.37	0.21	0.30	0.24	-0.07	0.00
ls-2	0.79	0.28	0.86	0.21	0.07	0.00
$mtsc_2$-2	0.52	0.37	0.57	0.35	0.05	0.09
sp-2	0.50	0.24	0.68	0.26	0.19	0.00
rsc-2	0.49	0.25	0.68	0.26	0.19	0.00
rsca-2	0.60	0.14	0.56	0.17	-0.04	0.00
avg_S-2	0.61	0.09	0.60	0.11	-0.01	0.20
maxstr-2	0.35	0.25	0.22	0.21	-0.13	0.00
dr-2	0.50	0.13	0.45	0.13	-0.05	0.00
si-2	0.44	0.20	0.37	0.24	-0.07	0.00
Φ_{ca}	-0.00	0.24	-0.01	0.28	-0.01	0.73
Φ_{ls}	0.20	0.39	0.16	0.34	-0.05	0.15
Φ_{mtsc_2}	0.15	0.39	0.12	0.39	-0.03	0.39
Φ_{sp}	0.12	0.25	0.24	0.33	0.11	0.00
Φ_{rsc}	0.12	0.26	0.23	0.33	0.11	0.00
Φ_{rsca}	-0.01	0.17	-0.01	0.25	0.01	0.74
Φ_{avg_S}	0.03	0.08	0.05	0.11	0.03	0.00
Φ_{maxstr}	-0.15	0.29	-0.19	0.30	-0.04	0.15
Φ_{dr}	-0.04	0.17	-0.05	0.17	-0.01	0.36

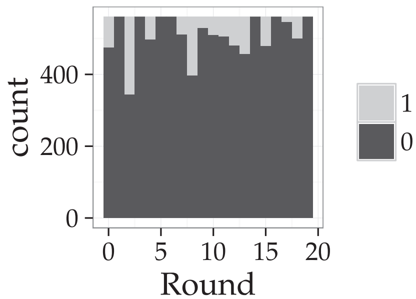




(a) Distribution of satisfaction during phase 1

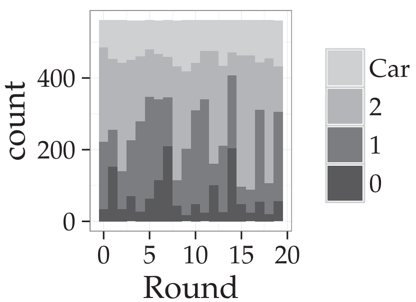
(b) Distribution of satisfaction during phase 2

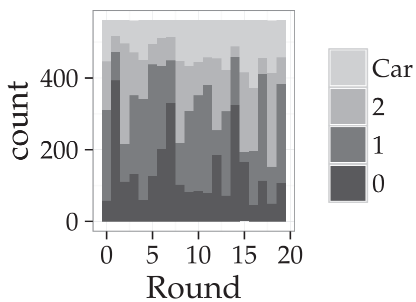




(c) Occurence of big disruptions during phase 1

(d) Occurence of big disruptions during phase 2





(e) Crowding levels experienced during phase 1

(f) Crowding levels experienced during phase 2

Figure B.1: Descriptive statistics of satisfaction, delays and crowding levels.

B.3 Numeric Data

Phase 1

$$
\mathrm{U} = \begin{bmatrix}
1 & 2 & 2 & 2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 2 & 1 & 1 & 2 & 2 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
1 & 2 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 0 & 2 & 1 & 2 & 2 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 0 & 2 & 1 & 1 & 1 \\
2 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 2 & 1 & 1 & 1 & 2 & 1 & 0 & 2 & 2 & 1 & 2 & 1 \\
1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 2 & 1 & 1 & 0 & 0 & 1 & 2
\end{bmatrix}
$$

$$
\mathrm{D} = \begin{bmatrix}
1 & 1 & 6 & 1 & 2 & 5 & 0 & 1 & 4 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 6 & 8 & 1 & 4 \\
4 & 3 & 4 & 1 & 2 & 0 & 1 & 2 & 7 & 6 & 1 & 5 & 3 & 1 & 1 & 3 & 3 & 5 & 0 & 0 \\
3 & 1 & 2 & 0 & 1 & 3 & 2 & 0 & 3 & 1 & 4 & 2 & 1 & 1 & 4 & 3 & 1 & 4 & 2 & 2 \\
2 & 3 & 1 & 1 & 2 & 3 & 3 & 2 & 5 & 1 & 4 & 3 & 0 & 4 & 7 & 0 & 4 & 4 & 3 & 1 \\
8 & 3 & 5 & 0 & 1 & 3 & 10 & 1 & 1 & 2 & 2 & 2 & 2 & 1 & 3 & 0 & 5 & 1 & 0 & 2 \\
1 & 1 & 3 & 3 & 1 & 1 & 2 & 3 & 5 & 3 & 3 & 1 & 1 & 1 & 4 & 0 & 2 & 0 & 3 & 2 \\
0 & 1 & 2 & 2 & 0 & 1 & 5 & 2 & 1 & 3 & 3 & 2 & 1 & 2 & 0 & 3 & 0 & 1 & 4 & 5
\end{bmatrix}
$$

$$
\mathrm{D}' = \begin{bmatrix}
1 & 1 & 6 & 1 & 2 & 5 & 0 & 1 & 4 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 6 & 8 & 1 & 4 \\
4 & 3 & 4 & 1 & 2 & 0 & 1 & 2 & 7 & 6 & 1 & 5 & 3 & 1 & 1 & 3 & 3 & 5 & 0 & 0 \\
3 & 1 & 2 & 0 & 1 & 3 & 2 & 0 & 3 & 1 & 4 & 2 & 1 & 1 & 4 & 3 & 1 & 4 & 2 & 2 \\
2 & 3 & 1 & 1 & 2 & 3 & 3 & 2 & 5 & 1 & 4 & 3 & 0 & 4 & 7 & 0 & 4 & 4 & 3 & 1 \\
8 & 3 & 5 & 0 & 1 & 3 & 10 & 1 & 1 & 2 & 2 & 2 & 2 & 1 & 3 & 0 & 5 & 1 & 0 & 2 \\
1 & 1 & 3 & 3 & 1 & 1 & 2 & 3 & 5 & 3 & 3 & 1 & 1 & 1 & 4 & 0 & 2 & 0 & 3 & 2 \\
10 & 3 & 1 & 7 & 6 & 1 & 6 & 0 & 7 & 5 & 12 & 2 & 1 & 5 & 2 & 3 & 2 & 7 & 10 & 1
\end{bmatrix}
$$

$$
\mathrm{DI} = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

Phase 2

$$
\mathrm{U} = \begin{bmatrix}
1 & 2 & 2 & 2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 0 & 2 & 1 & 1 & 2 & 2 & 1 & 2 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\
1 & 2 & 2 & 2 & 1 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 0 & 2 & 1 & 2 & 2 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 0 & 2 & 1 & 1 & 1 \\
2 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 2 & 1 & 1 & 1 & 2 & 1 & 0 & 2 & 2 & 1 & 2 & 1 \\
1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 2 & 1 & 1 & 0 & 0 & 1 & 2
\end{bmatrix}
$$

$$
\mathrm{D} = \begin{bmatrix}
1 & 1 & 6 & 1 & 2 & 5 & 0 & 1 & 4 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 6 & 8 & 1 & 4 \\
4 & 3 & 4 & 1 & 2 & 0 & 1 & 2 & 7 & 6 & 1 & 5 & 3 & 1 & 1 & 3 & 3 & 5 & 0 & 0 \\
3 & 1 & 2 & 0 & 1 & 3 & 2 & 0 & 3 & 1 & 4 & 2 & 1 & 1 & 4 & 3 & 1 & 4 & 2 & 2 \\
2 & 3 & 1 & 1 & 2 & 3 & 3 & 2 & 5 & 1 & 4 & 3 & 0 & 4 & 7 & 0 & 4 & 4 & 3 & 1 \\
8 & 3 & 5 & 0 & 1 & 3 & 10 & 1 & 1 & 2 & 2 & 2 & 2 & 1 & 3 & 0 & 5 & 1 & 0 & 2 \\
1 & 1 & 3 & 3 & 1 & 1 & 2 & 3 & 5 & 3 & 3 & 1 & 1 & 1 & 4 & 0 & 2 & 0 & 3 & 2 \\
0 & 1 & 2 & 2 & 0 & 1 & 5 & 2 & 1 & 3 & 3 & 2 & 1 & 2 & 0 & 3 & 0 & 1 & 4 & 5
\end{bmatrix}
$$

$$D' = \begin{bmatrix} 1 & 1 & 6 & 1 & 2 & 5 & 0 & 1 & 4 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 6 & 8 & 1 & 4 \\ 4 & 3 & 4 & 1 & 2 & 0 & 1 & 2 & 7 & 6 & 1 & 5 & 3 & 1 & 1 & 3 & 3 & 5 & 0 & 0 \\ 3 & 1 & 2 & 0 & 1 & 3 & 2 & 0 & 3 & 1 & 4 & 2 & 1 & 1 & 4 & 3 & 1 & 4 & 2 & 2 \\ 2 & 3 & 1 & 1 & 2 & 3 & 3 & 2 & 5 & 1 & 4 & 3 & 0 & 4 & 7 & 0 & 4 & 4 & 3 & 1 \\ 8 & 3 & 5 & 0 & 1 & 3 & 10 & 1 & 1 & 2 & 2 & 2 & 2 & 1 & 3 & 0 & 5 & 1 & 0 & 2 \\ 1 & 1 & 3 & 3 & 1 & 1 & 2 & 3 & 5 & 3 & 3 & 1 & 1 & 1 & 4 & 0 & 2 & 0 & 3 & 2 \\ 10 & 3 & 1 & 7 & 6 & 1 & 6 & 0 & 7 & 5 & 12 & 2 & 1 & 5 & 2 & 3 & 2 & 7 & 10 & 1 \end{bmatrix}$$

$$DI = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$I = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 1 & 0 & 1 & 1 \\ 2 & 2 & 2 & 1 & 2 & 1 & 0 & 0 & 2 & 1 & 0 & 2 & 0 & 1 & 1 & 1 & 2 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 1 & 2 & 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 2 & 1 & 0 & 1 & 1 & 2 & 0 & 2 & 1 & 1 \\ 2 & 1 & 1 & 1 & 2 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 1 \\ 2 & 0 & 1 & 0 & 2 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 2 & 1 & 2 & 2 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Back Matter

English Summary

Peak demands are a costly affair for public transport operators as they dictate the required infrastructure and vehicle capacity. As a consequence vehicles are crowded during a specific part of the day, while some vehicles remain unutilized for the remainder of the time. In this thesis we study how microscopic data can be used in models that have the potential to improve utilization, while preventing excess crowding. We performed this study in three parts. First, we model crowding behavior due to interactions between the behavior of passengers and the public transport operators who plan the vehicle capacities. Second, we develop algorithms that can extract temporal and spatial patterns from smart card data. Finally, we consider how we can develop individual decision strategies in situations where there is uncertainty about travel options.

In the first part we introduce a simulation model in which individual travelers choose repeatedly at which time they want to travel. Their aim is to end up in a vehicle that is not too crowded, or to travel by different means if no such travel option is available. In a first experiment we analyze how the available amount of information an operator discloses influences the utilization of vehicles and the satisfaction of artificial passengers. In a second experiment we analyze the interactions between repeated travel choices of such passengers and the planning of capacities performed by a public transport operator on a single line with a number of stations. In addition to simulation experiments we performed a quasi-experiment with a large number of students in a similar setting as the simulation experiments. We found four behavioral groups: Stoics who stick to a single travel choice every step of the experiment, dualists who switch between two preferred options, moderate switchers who keep switching between travel options throughout the experiment and heavy switchers who almost never stick to their previous choice. The introduction of information related to crowding influences the sizes of these behavioral groups.

In the second part we make a first step to convert smart card data to input for the agent-based transport simulation software MATSim. We also develop algorithms that look for frequently occurring activity patterns in a smart card data set. Finally, we propose a way to generate synthetic smart card data. This data allows us to compare the patterns used to generate the data to the patterns found by an algorithm that analyzes such data. This is useful for the empirical validation of such algorithms.

In the third part we consider how to make travel decisions under uncertainty, for example about the duration of a disruption. We introduce the *traveler's route choice* problem, in which an individual traveler has to decide between two periodically departing travel options. One of the alternatives is unavailable for an unknown amount of time. First we investigate which strategies are best for a specific type of objective (the competitive ratio) and then we analyze which strategies are worthwhile to consider for a very broad set of objectives.

Nederlandse Samenvatting

De spits is een kostbaar fenomeen voor openbaarvervoerbedrijven, omdat de benodigde infrastructuur en voertuigcapaciteiten door de vraag in de spits bepaald worden. Als gevolg hiervan ontstaat op specifieke momenten grote drukte en blijft een deel van de voertuigcapaciteit de rest van de dag onbenut. In dit proefschrift onderzoeken we de mogelijke rol van microscopische gegevens in modellen die potentie hebben de bezettingsgraad van het openbaar vervoer te verbeteren, zonder extreme druktes te introduceren. We hebben dit onderzoek in drie delen uitgevoerd. Ten eerste modelleren we het ontstaan van overvolle voertuigen als gevolg van interacties tussen het keuzegedrag van reizigers en de capaciteitsplanning van openbaarvervoerbedrijven. Ten tweede ontwikkelen we algoritmen die tijd- en ruimtepatronen kunnen destilleren uit chipkaartgegevens. Ten slotte beschouwen we hoe individuele beslisstrategieën voor situaties met onzekerheid over reisopties ontwikkeld kunnen worden.

In het eerste deel introduceren we een simulatiemodel waarbinnen individuele reizigers herhaaldelijk kiezen op welk tijdstip ze willen reizen. Hun voorkeur is te reizen met een voertuig dat niet te vol is en wanneer dit niet kan op een andere manier te reizen. In een eerste experiment analyseren we hoe de beschikbaarheid van informatie die door de vervoerder wordt vrijgegeven, de bezettingsgraad van de voertuigen en de tevredenheid van kunstmatige passagiers beïnvloedt. In een tweede experiment analyseren we de interactie tussen herhaalde keuzes van de reizigers en de capaciteitsplanning door de vervoerder op een lijn met enkele stations. Naast deze simulaties hebben we een quasi-experiment uitgevoerd met een grote groep studenten in een met de simulaties vergelijkbare opzet. Er zijn vier verschillende typen gedrag zichtbaar: stoïcijnen blijven altijd bij hun eerste keus, dualisten blijven wisselen tussen twee voorkeursopties, gematigde wisselaars die meerdere reisopties blijven overwegen en als variant van die groep de frequente wisselaars. Door informatie over drukte te introduceren tijdens het experiment worden de groepsgrootten van deze gedragstypen beïnvloed.

In het tweede deel doen we een eerste stap om chipkaartgegevens als invoer van het agent-based simulatiepakket MATSim te gebruiken. We ontwikkelen hiervoor algoritmen die zoeken naar frequent terugkerende activiteitenpatronen in zulke gegevens. Ten slotte ontwikkelen we een methode om synthetische chipkaartgegevens te genereren. Met deze gegevens kunnen we patronen, die gebruikt zijn om de data te genereren, vergelijken met de patronen die worden gevonden door algoritmen voor data-analyse. Dit is nuttig voor empirische validatie van dergelijke algoritmen.

In het derde deel beschouwen we het maken van reisbeslissingen bij onzekerheid over bijvoorbeeld de duur van een verstoring. Hiervoor introduceren we het *traveler's route choice* probleem, waarbinnen een individuele reiziger kiest tussen twee periodiek vertrekkende reisalternatieven. Hier is één van de alternatieven onbeschikbaar voor een onbekende tijdsduur. Om te beginnen analyseren we welke strategieën het beste zijn met het oog op een specifieke doelstelling (de "competitive ratio"). Ten slotte analyseren we welke strategieën überhaupt het overwegen waard zijn voor een breed scala aan mogelijke doelstellingen.

About the Author

Paul Bouman was born in the year 1984 in Bilthoven, The Netherlands. He obtained a cum laude Bachelor of Computer Science and a cum laude Master of Applied Computer Science at Utrecht University. During his bachelor studies he obtained a minor in Artificial Intelligence and programmed robotic dogs playing soccer in the AIBO League of the RoboCup. His master thesis was on "A column generation framework for recoverable robustness".

In 2011, Paul started his PhD research at the Rotterdam School of Management, Erasmus University within the Complexity in Public Transport (ComPuTr) project of the Netherlands Organisation for Scientific Research (NWO). As part of this project Paul focused on the modeling of peak demands in relation to public transport optimization. In his research he focuses on the design and implementation of algorithms, models for scheduling and optimization, decision making in uncertain situations, simulation of (complex) systems and the processing of large data sets. In most of his research these techniques are applied to the domain of transportation in general and public transport in particular.
Paul attended multiple international conferences to present his research, including EURO, IFORS, TRISTAN and others. Currently, Paul works as an Assistant Professor within the Econometric Institute of the Erasmus School of Economics, where he conducts research and teaching related to Computer Science and Operations Research.

Author Portfolio

Journal Articles

P.C. Bouman, M. Schmidt, L.G. Kroon, and A. Schobel (2013, October). Passenger route choice in case of disruptions. *In Intelligent Transportation Systems-(ITSC), 2013 16th International IEEE Conference on (pp. 1047-1052)*. IEEE.

J.M. van den Akker, **P.C. Bouman**, J.A. Hoogeveen, and D.D. Tönissen (2016). Decomposition approaches for recoverable robust optimization problems. *European Journal of Operational Research, 251*(3), 739-750.

P.C. Bouman, L.G. Kroon, P.H.M. Vervest, and G. Maróti (2016). Capacity, information and minority games in public transport. *Transportation Research Part C: Emerging Technologies, 70*, 157-170.

M. Schmidt, L.G. Kroon, A. Schöbel, and **P.C. Bouman** (2017). The Travelers Route Choice Problem Under Uncertainty: Dominance Relations Between Strategies. *Operations Research*.

Peer Reviewed Conference Proceedings

P.C. Bouman, J.M. van den Akker, and J.A. Hoogeveen (2011, September). Recoverable robustness by column generation. In *European Symposium on Algorithms* (pp. 215-226). Springer Berlin Heidelberg.

P.C. Bouman, M. Lovric, T. Li, E. van der Hurk, L.G. Kroon and P.H.M. Vervest (2012). Recognizing Demand Patterns from Smart Card Data for Agent-Based Microsimulation of Public Transport. In *ATT 2012: 7th International Workshop on Agents in Traffic and Transportation, June 5th, Valencia, Spain*.

P.C. Bouman, E. van der Hurk, L.G. Kroon, T. Li, and P.H.M. Vervest (2013, November). Detecting activity patterns from smart card Data. In *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence, Delft, The Netherlands, November 7-8, 2013*.

Submitted Papers

N.A.H. Agatz, **P.C. Bouman** and M. Schmidt (2016) Optimization Approaches for the Traveling Salesman Problem with Drone. Under major revision at *Transportation Research*.

Working Papers

P.C. Bouman, C.L.P. Pennings, J. van Dalen and L.G. Kroon. Time Choice Data for Public Transport Optimization.

K.M. Glorie, M. Carvalho, M. Constantino, **P.C. Bouman** and A. Viana. Robust Models for the Kidney Exchange Program.

Other Work

P.C. Bouman, S. Selaković and M. van der Leij (2012). Modularity and Innovation. *Assignment of the NWO Complexity Winter School, December 2011*.

P.C. Bouman, E. Siero and D. in 't Veld (2013). Second line of Defense in Cyber Security: Intelligent Allocation of Virtual Machines. *Assignment of the NWO Complexity Winter School, December 2012*.

P.C. Bouman and M. Lovric (2016). Rotterdam: Revenue Management in Public Transportation with Smart-Card Data Enabled Agent-Based Simulations. In A. Horni, K. Nagel and K.W. Axhausen (eds.) *The Multi-Agent Transport Simulation MATSim*. London: Ubiquity Press. (Book chapter)

PhD Courses

Algorithmic Game Theory
CO1a: Networks and Polyhedra
CO1b: Networks and Semidefinite Programming
CO2a: Algorithms & Complexity
CO2b: Integer Programming Methods
Convex Analysis for Optimization
Cooperative Games
Interior Point Methods
Noncooperative Games
OR-Games
Publishing Strategy
Randomized Algorithms
Robust Optimization
Statistical Methods
Stochastic Models and Optimisation
Stochastic Programming

Teaching

As Main Lecturer and Coordinator

Programming
Advanced Programming
ICT

As Tutorial Lecturer

Big Data and Business Analytics
Quantitative Methods & Techniques: Mathematics
Decision Support

As Student Assistant

Imperative programming
Computer architecture and networks
Databases
Graphics
Search algorithms
Data structures
Mathematical techniques for computer science
Logic & Set Theory
Expert Systems
Functional programming
Internet Programming
Distributed Programming
Strategic Management and ICT

Conferences

2011	**BNAIC Conference**	Ghent, Belgium
2012	**LNMB Conference**	Lunteren, The Netherlands
2012	**EURO Conference**	Vilnius, Lithuania
2012	**AAMAS Conference**	Valencia, Spain
2012	**CASPT Conference**	Santiago, Chile
2012	**BNAIC Conference**	Maastricht, The Netherlands
2012	**MATSim User Meeting**	Berlin, Germany
2013	**LNMB Conference**	Lunteren, The Netherlands
2013	**RailCopenhagen Conference**	Copenhagen, Denmark
2013	**TRISTAN Conference**	San Pedro de Atacama, Chile
2013	**OR Conference**	Rotterdam, The Netherlands
2013	**IEEE-ITSC Conference**	The Hague, The Netherlands
2013	**BNAIC Conference**	Delft, The Netherlands
2014	**LNMB Conference**	Lunteren, The Netherlands
2014	**AAMAS Conference**	Paris, France

2014	**IFORS Conference**	Barcelona, Spain
2015	**LNMB Conference**	Lunteren, The Netherlands
2015	**RailTokyo Conference**	Tokyo, Japan
2015	**TSL Workshop**	Berlin, Germany
2015	**CASPT Conference**	Rotterdam, The Netherlands
2016	**LNMB Conference**	Lunteren, The Netherlands
2016	**TRISTAN Conference**	Oranjestad, Aruba
2017	**LNMB Conference**	Lunteren, The Netherlands

Committee Memberships

Organizing Committee CASPT 2015, Rotterdam, The Netherlands.
Scientific Review Committee TRISTAN IX 2016, Oranjestad, Aruba.

The ERIM PhD Series

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

Dissertations in the last five years

Abbink, E.J., *Crew Management in Passenger Rail Transport*, Promotors: Prof. L.G. Kroon & Prof. A.P.M. Wagelmans, EPS-2014-325-LIS, http://repub.eur.nl/pub/76927

Acar, O.A., *Crowdsourcing for Innovation: Unpacking Motivational, Knowledge and Relational Mechanisms of Innovative Behavior in Crowdsourcing Platforms*, Promotor: Prof. J.C.M. van den Ende, EPS-2014-321-LIS, http://repub.eur.nl/pub/76076

Akin Ates, M., *Purchasing and Supply Management at the Purchase Category Level: Strategy, structure and performance*, Promotors: Prof. J.Y.F. Wynstra & Dr E.M. van Raaij, EPS-2014-300-LIS, http://repub.eur.nl/pub/50283

Akpinar, E., *Consumer Information Sharing*, Promotor: Prof. A. Smidts, EPS-2013-297-MKT, http://repub.eur.nl/pub/50140

Alexander, L., *People, Politics, and Innovation: A Process Perspective*, Promotors: Prof. H.G. Barkema & Prof. D.L. van Knippenberg, EPS-2014-331-S&E, http://repub.eur.nl/pub/77209

Alexiou, A., *Management of Emerging Technologies and the Learning Organization: Lessons from the Cloud and Serious Games Technology*, Promotor: Prof. S.J. Magala, Prof. M.C. Schippers and Dr I. Oshri, EPS-2016-404-ORG, http://repub.eur.nl/pub/93818

Almeida e Santos Nogueira, R.J. de, *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*, Promotors: Prof. U. Kaymak & Prof. J.M.C. Sousa, EPS-2014-310-LIS, http://repub.eur.nl/pub/51560

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-frequency Data*, Promotor: Prof. D.J.C. van Dijk, EPS-2013-273-F&A, http://repub.eur.nl/pub/38240

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promotors: Prof. H.W. Volberda & Prof. F.A.J. van den Bosch, EPS-2013-278-S&E, http://repub.eur.nl/pub/39128

Benschop, N., *Biases in Project Escalation: Names, frames & construal levels*, Promotors: Prof. K.I.M. Rhode, Prof. H.R. Commandeur, Prof. M. Keil & Dr A.L.P. Nuijten, EPS-2015-375-S&E, http://repub.eur.nl/pub/79408

Berg, W.E. van den, *Understanding Salesforce Behavior using Genetic Association Studies*, Promotor: Prof. W.J.M.I. Verbeke, EPS-2014-311-MKT, http://repub.eur.nl/pub/51440

Beusichem, H.C. van, *Firms and Financial Markets: Empirical Studies on the Informational Value of Dividends, Governance and Financial Reporting*, Promotors: Prof. A. de Jong & Dr G. Westerhuis, EPS-2016-378-F&A, http://repub.eur.nl/pub/93079

Bliek, R. de, *Empirical Studies on the Economic Impact of Trust*, Promotors: Prof. J. Veenman & Prof. Ph.H.B.F. Franses, EPS-2015-324-ORG, http://repub.eur.nl/pub/78159

Boons, M., *Working Together Alone in the Online Crowd: The Effects of Social Motivations and Individual Knowledge Backgrounds on the Participation and Performance of Members of Online Crowdsourcing Platforms*, Promotors: Prof. H.G. Barkema & Dr D.A. Stam, EPS-2014-306-S&E, http://repub.eur.nl/pub/50711

Brazys, J., *Aggregated Marcoeconomic News and Price Discovery*, Promotor: Prof. W.F.C. Verschoor, EPS-2015-351-F&A, http://repub.eur.nl/pub/78243

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit, Coworker Satisfaction, and Relational Models*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-292-ORG, http://repub.eur.nl/pub/41508

Cancurtaran, P., *Essays on Accelerated Product Development*, Promotors: Prof. F. Langerak & Prof. G.H. van Bruggen, EPS-2014-317-MKT, http://repub.eur.nl/pub/76074

Caron, E.A.M., *Explanation of Exceptional Values in Multi-dimensional Business Databases*, Promotors: Prof. H.A.M. Daniels & Prof. G.W.J. Hendrikse, EPS-2013-296-LIS, http://repub.eur.nl/pub/50005

Carvalho, L. de, *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promotor: Prof. L. Berg, EPS-2013-274-S&E, http://repub.eur.nl/pub/38449

Cranenburgh, K.C. van, *Money or Ethics: Multinational corporations and religious organisations operating in an era of corporate responsibility*, Promotors: Prof. L.C.P.M. Meijs, Prof. R.J.M. van Tulder & Dr D. Arenas, EPS-2016-385-ORG, http://repub.eur.nl/pub/93104

Consiglio, I., *Others: Essays on Interpersonal and Consumer Behavior*, Promotor: Prof. S.M.J. van Osselaer, EPS-2016-366-MKT, http://repub.eur.nl/pub/79820

Cox, R.H.G.M., *To Own, To Finance, and To Insure - Residential Real Estate Revealed*, Promotor: Prof. D. Brounen, EPS-2013-290-F&A, http://repub.eur.nl/pub/40964

Darnihamedani, P., *Individual Characteristics, Contextual Factors and Entrepreneurial Behavior*, Promotors: Prof. A.R. Thurik & S.J.A. Hessels, EPS-2016-360-S&E, http://repub.eur.nl/pub/93280

Deng, W., *Social Capital and Diversification of Cooperatives*, Promotor: Prof. G.W.J. Hendrikse, EPS-2015-341-ORG, http://repub.eur.nl/pub/77449

Depecik, B.E., *Revitalizing brands and brand: Essays on Brand and Brand Portfolio Management Strategies*, Promotor: Prof. G.H. van Bruggen, Dr Y.M. van Everdingen and Dr M.B. Ataman, EPS-2016-406-MKT, http://repub.eur.nl/pub/93507

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promotor: Prof. A.P.M. Wagelmans, EPS-2013-272-LIS, http://repub.eur.nl/pub/38241

Duyvesteyn, J.G., *Empirical Studies on Sovereign Fixed Income Markets*, Promotors: Prof. P. Verwijmeren & Prof. M.P.E. Martens, EPS-2015-361-F&A, http://hdl.handle.net/1765/79033

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-Follower Dyad*, Promotor: Prof. R.J.M. van Tulder, EPS-2013-279-ORG, http://repub.eur.nl/pub/39129

Elemes, A., *Studies on Determinants and Consequences of Financial Reporting Quality*, Promotor: Prof. E. Peek, EPS-2015-354-F&A, http://hdl.handle.net/1765/79037

Ellen, S. ter, *Measurement, Dynamics, and Implications of Heterogeneous Beliefs in Financial Markets*, Promotor: Prof. W.F.C. Verschoor, EPS-2015-343-F&A, http://repub.eur.nl/pub/78191

Erlemann, C., *Gender and Leadership Aspiration: The Impact of the Organizational Environment*, Promotor: Prof. D.L. van Knippenberg, EPS-2016-376-ORG, http://repub.eur.nl/pub/79409

Eskenazi, P.I., *The Accountable Animal*, Promotor: Prof. F.G.H. Hartmann, EPS- 2015-355-F&A, http://repub.eur.nl/pub/78300

Evangelidis, I., *Preference Construction under Prominence*, Promotor: Prof. S.M.J. van Osselaer, EPS-2015-340-MKT, http://repub.eur.nl/pub/78202

Faber, N., *Structuring Warehouse Management*, Promotors: Prof. M.B.M. de Koster & Prof. A. Smidts, EPS-2015-336-LIS, http://repub.eur.nl/pub/78603

Fernald, K., *The Waves of Biotechnological Innovation in Medicine: Interfirm Cooperation Effects and a Venture Capital Perspective*, Promotors: Prof. E. Claassen, Prof. H.P.G. Pennings & Prof. H.R. Commandeur, EPS-2015-371-S&E, http://hdl.handle.net/1765/79120

Fisch, C.O., *Patents and trademarks: Motivations, antecedents, and value in industrialized and emerging markets*, Promotors: Prof. J.H. Block, Prof. H.P.G. Pennings & Prof. A.R. Thurik, EPS-2016-397-S&E, http://repub.eur.nl/pub/94036

Fliers, P.T., *Essays on Financing and Performance: The role of firms, banks and board*, Promotors: Prof. A. de Jong & Prof. P.G.J. Roosenboom, EPS-2016-388-F&A, http://repub.eur.nl/pub/93019

Fourne, S.P., *Managing Organizational Tensions: A Multi-Level Perspective on Exploration, Exploitation and Ambidexterity*, Promotors: Prof. J.J.P. Jansen & Prof. S.J. Magala, EPS-2014-318-S&E, http://repub.eur.nl/pub/76075

Gaast, J.P. van der, *Stochastic Models for Order Picking Systems*, Promotors: Prof. M.B.M de Koster & Prof. I.J.B.F. Adan, EPS-2016-398-LIS, http://repub.eur.nl/pub/93222

Glorie, K.M., *Clearing Barter Exchange Markets: Kidney Exchange and Beyond*, Promotors: Prof. A.P.M. Wagelmans & Prof. J.J. van de Klundert, EPS-2014-329-LIS, http://repub.eur.nl/pub/77183

Hekimoglu, M., *Spare Parts Management of Aging Capital Products*, Promotor: Prof. R. Dekker, EPS-2015-368-LIS, http://repub.eur.nl/pub/79092

Heyde Fernandes, D. von der, *The Functions and Dysfunctions of Reminders*, Promotor: Prof. S.M.J. van Osselaer, EPS-2013-295-MKT, http://repub.eur.nl/pub/41514

Hogenboom, A.C., *Sentiment Analysis of Text Guided by Semantics and Structure*, Promotors: Prof. U. Kaymak & Prof. F.M.G. de Jong, EPS-2015-369-LIS, http://repub.eur.nl/pub/79034

Hogenboom, F.P., *Automated Detection of Financial Events in News Text*, Promotors: Prof. U. Kaymak & Prof. F.M.G. de Jong, EPS-2014-326-LIS, http://repub.eur.nl/pub/77237

Hollen, R.M.A., *Exploratory Studies into Strategies to Enhance Innovation-Driven International Competitiveness in a Port Context: Toward Ambidextrous Ports*, Promotors: Prof. F.A.J. Van Den Bosch & Prof. H.W.Volberda, EPS-2015-372-S&E, http://repub.eur.nl/pub/78881

Hout, D.H. van, *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*, Promotors: Prof. P.J.F. Groenen & Prof. G.B. Dijksterhuis, EPS- 2014-304-MKT, `http://repub.eur.nl/pub/50387`

Houwelingen, G.G. van, *Something To Rely On*, Promotors: Prof. D. de Cremer & Prof. M.H. van Dijke, EPS-2014-335-ORG, `http://repub.eur.nl/pub/77320`

Hurk, E. van der, *Passengers, Information, and Disruptions*, Promotors: Prof. L.G. Kroon & Prof. P.H.M. Vervest, EPS-2015-345-LIS, `http://repub.eur.nl/pub/78275`

Iseger, P. den, *Fourier and Laplace Transform Inversion with Applications in Finance*, Promotor: Prof. R. Dekker, EPS-2014-322-LIS, `http://repub.eur.nl/pub/76954`

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promotor: Prof. R. Dekker, EPS-2013-288-LIS, `http://repub.eur.nl/pub/39933`

Khanagha, S., *Dynamic Capabilities for Managing Emerging Technologies*, Promotor: Prof. H.W. Volberda, EPS-2014-339-S&E, `http://repub.eur.nl/pub/77319`

Kil, J., *Acquisitions Through a Behavioral and Real Options Lens*, Promotor: Prof. H.T.J. Smit, EPS-2013-298-F&A, `http://repub.eur.nl/pub/50142`

Klooster, E. van't, *Travel to Learn: the Influence of Cultural Distance on Competence Development in Educational Travel*, Promotors: Prof. F.M. Go & Prof. P.J. van Baalen, EPS-2014-312-MKT, `http://repub.eur.nl/pub/51462`

Koendjbiharie, S.R., *The Information-Based View on Business Network Performance: Revealing the Performance of Interorganizational Networks*, Promotors: Prof. H.W.G.M. van Heck & Prof. P.H.M. Vervest, EPS-2014-315-LIS, `http://repub.eur.nl/pub/51751`

Koning, M., *The Financial Reporting Environment: The Role of the Media, Regulators and Auditors*, Promotors: Prof. G.M.H. Mertens & Prof. P.G.J. Roosenboom, EPS-2014-330-F&A, `http://repub.eur.nl/pub/77154`

Konter, D.J., *Crossing Borders with HRM: An Inquiry of the Influence of Contextual Differences in the Adoption and Effectiveness of HRM*, Promotors: Prof. J. Paauwe & Dr L.H. Hoeksema, EPS-2014-305-ORG, `http://repub.eur.nl/pub/50388`

Korkmaz, E., *Bridging Models and Business: Understanding Heterogeneity in Hidden Drivers of Customer Purchase Behavior*, Promotors: Prof. S.L. van de Velde & Prof. D. Fok, EPS-2014-316-LIS, `http://repub.eur.nl/pub/76008`

Krämer, R., *A license to mine? Community organizing against multinational corporations*, Promotors: Prof. R.J.M. van Tulder & Prof. G.M. Whiteman, EPS-2016-383-ORG, http://repub.eur.nl/pub/94072

Kroezen, J.J., *The Renewal of Mature Industries: An Examination of the Revival of the Dutch Beer Brewing Industry*, Promotor: Prof. P.P.M.A.R. Heugens, EPS-2014- 333-S&E, http://repub.eur.nl/pub/77042

Kysucky, V., *Access to Finance in a Cros-Country Context*, Promotor: Prof. L. Norden, EPS-2015-350-F&A, http://repub.eur.nl/pub/78225

Lee, C.I.S.G., *Big Data in Management Research: Exploring New Avenues*, Promotors: Prof. S.J. Magala & Dr W.A. Felps, EPS-2016-365-ORG, http://repub.eur.nl/pub/79818

Legault-Tremblay, P.O., *Corporate Governance During Market Transition: Heterogeneous responses to Institution Tensions in China*, Promotor: Prof. B. Krug, EPS-2015-362-ORG, http://repub.eur.nl/pub/78649

Lenoir, A.S., *Are You Talking to Me? Addressing Consumers in a Globalised World*, Promotors: Prof. S. Puntoni & Prof. S.M.J. van Osselaer, EPS-2015-363-MKT, http://repub.eur.nl/pub/79036

Leunissen, J.M., *All Apologies: On the Willingness of Perpetrators to Apologize*, Promotors: Prof. D. de Cremer & Dr M. van Dijke, EPS-2014-301-ORG, http://repub.eur.nl/pub/50318

Li, D., *Supply Chain Contracting for After-sales Service and Product Support*, Promotor: Prof. M.B.M. de Koster, EPS-2015-347-LIS, http://repub.eur.nl/pub/78526

Li, Z., *Irrationality: What, Why and How*, Promotors: Prof. H. Bleichrodt, Prof. P.P. Wakker, & Prof. K.I.M. Rohde, EPS-2014-338-MKT, http://repub.eur.nl/pub/77205

Liang, Q.X., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives*, Promotor: Prof. G.W.J. Hendrikse, EPS-2013-281-ORG, http://repub.eur.nl/pub/39253

Liket, K., *Why 'Doing Good' is not Good Enough: Essays on Social Impact Measurement*, Promotors: Prof. H.R. Commandeur & Dr K.E.H. Maas, EPS-2014-307-STR, http://repub.eur.nl/pub/51130

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research*, Promotors: Prof. A.R. Thurik, Prof. P.J.F. Groenen, & Prof. A. Hofman, EPS-2013-287-S&E, http://repub.eur.nl/pub/40081

Lu, Y., *Data-Driven Decision Making in Auction Markets*, Promotors: Prof. H.W.G.M. van Heck & Prof. W. Ketter, EPS-2014-314-LIS, http://repub.eur.nl/pub/51543

Ma, Y., *The Use of Advanced Transportation Monitoring Data for Official Statistics*, Promotor: Prof. L.G. Kroon and Dr J. van Dalen, EPS-2016-391-LIS, http://repub.eur.nl/pub/80174

Manders, B., *Implementation and Impact of ISO 9001*, Promotor: Prof. K. Blind, EPS-2014-337-LIS, http://repub.eur.nl/pub/77412

Mell, J.N., *Connecting Minds: On The Role of Metaknowledge in Knowledge Coordination*, Promotor: Prof. D.L. van Knippenberg, EPS-2015-359-ORG, http://hdl.handle.net/1765/78951

Meulen, D. van der, *The Distance Dilemma: the effect of flexible working practices on performance in the digital workplace*, Promotors: Prof. H.W.G.M. van Heck & Prof. P.J. van Baalen, EPS-2016-403-LIS, http://repub.eur.nl/pub/94033

Micheli, M.R., *Business Model Innovation: A Journey across Managers' Attention and Inter- Organizational Networks*, Promotor: Prof. J.J.P. Jansen, EPS-2015-344-S&E, http://repub.eur.nl/pub/78241

Milea, V., *News Analytics for Financial Decision Support*, Promotor: Prof. U. Kaymak, EPS-2013-275-LIS, http://repub.eur.nl/pub/38673

Moniz, A., *Textual Analysis of Intangible Information*, Promotors: Prof. C.B.M. van Riel, Prof. F.M.G de Jong & Dr G.A.J.M. Berens, EPS-2016-393-ORG, http://repub.eur.nl/pub/93001

Mulder, J., *Network design and robust scheduling in liner shipping*, Promotors: Prof. R. Dekker & Dr W.L. van Jaarsveld, EPS-2016-384-LIS, http://repub.eur.nl/pub/80258

Naumovska, I., *Socially Situated Financial Markets: A Neo-Behavioral Perspective on Firms, Investors and Practices*, Promotors: Prof. P.P.M.A.R. Heugens & Prof. A. de Jong, EPS-2014-319-S&E, http://repub.eur.nl/pub/76084

Neerijnen, P., *The Adaptive Organization: the socio-cognitive antecedents of ambidexterity and individual exploration*, Promotors: Prof. J.J.P. Jansen, P.P.M.A.R. Heugens & Dr T.J.M. Mom, EPS-2016-358-S&E, http://repub.eur.nl/pub/93274

Oord, J.A. van, *Essays on Momentum Strategies in Finance*, Promotor: Prof. H.K. van Dijk, EPS-2016-380-F&A, http://repub.eur.nl/pub/80036

Pennings, C.L.P., *Advancements in Demand Forecasting: Methods and Behavior*, Promotors: Prof. L.G. Kroon, Prof. H.W.G.M. van Heck & Dr J. van Dalen, EPS-2016-400-LIS, http://repub.eur.nl/pub/94039

Peters, M., *Machine Learning Algorithms for Smart Electricity Markets*, Promotor: Prof. W. Ketter, EPS-2014-332-LIS, http://repub.eur.nl/pub/77413

Porck, J., *No Team is an Island: An Integrative View of Strategic Consensus between Groups*, Promotors: Prof. P.J.F. Groenen & Prof. D.L. van Knippenberg, EPS- 2013-299-ORG, http://repub.eur.nl/pub/50141

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promotors: Prof. H.J.H.M. Claassen & Prof. H.R. Commandeur, EPS-2013-282-S&E, http://repub.eur.nl/pub/39654

Protzner, S., *Mind the gap between demand and supply: A behavioral perspective on demand forecasting*, Promotors: Prof. S.L. van de Velde & Dr L. Rook, EPS-2015-364-LIS, http://repub.eur.nl/pub/79355

Pruijssers, J.K., *An Organizational Perspective on Auditor Conduct*, Promotors: Prof. J. van Oosterhout & Prof. P.P.M.A.R. Heugens, EPS-2015-342-S&E, http://repub.eur.nl/pub/78192

Retel Helmrich, M.J., *Green Lot-Sizing*, Promotor: Prof. A.P.M. Wagelmans, EPS- 2013-291-LIS, http://repub.eur.nl/pub/41330

Rietdijk, W.J.R., *The Use of Cognitive Factors for Explaining Entrepreneurship*, Promotors: Prof. A.R. Thurik & Prof. I.H.A. Franken, EPS-2015-356-S&E, http://repub.eur.nl/pub/79817

Rietveld, N., *Essays on the Intersection of Economics and Biology*, Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger, Prof. P.J.F. Groenen, & Prof. A. Hofman, EPS-2014-320-S&E, http://repub.eur.nl/pub/76907

Rösch, D., *Market Efficiency and Liquidity*, Promotor: Prof. M.A. van Dijk, EPS-2015-353-F&A, http://repub.eur.nl/pub/79121

Roza, L., *Employee Engagement in Corporate Social Responsibility: A collection of essays*, Promotor: L.C.P.M. Meijs, EPS-2016-396-ORG, http://repub.eur.nl/pub/93254

Rubbaniy, G., *Investment Behaviour of Institutional Investors*, Promotor: Prof. W.F.C. Verschoor, EPS-2013-284-F&A, http://repub.eur.nl/pub/40068

Schoonees, P., *Methods for Modelling Response Styles*, Promotor: Prof.dr P.J.F. Groenen, EPS-2015-348-MKT, http://repub.eur.nl/pub/79327

Schouten, M.E., *The Ups and Downs of Hierarchy: the causes and consequences of hierarchy struggles and positional loss*, Promotors: Prof. D.L. van Knippenberg & Dr L.L. Greer, EPS-2016-386-ORG, http://repub.eur.nl/pub/80059

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promotor: Prof. G.M.H. Mertens, EPS-2013-283-F&A, http://repub.eur.nl/pub/39655

Smit, J., *Unlocking Business Model Innovation: A look through the keyhole at the inner workings of Business Model Innovation*, Promotor: H.G. Barkema, EPS-2016-399-S&E, http://repub.eur.nl/pub/93211

Sousa, M.J.C. de, *Servant Leadership to the Test: New Perspectives and Insights*, Promotors: Prof. D.L. van Knippenberg & Dr D. van Dierendonck, EPS-2014-313-ORG, http://repub.eur.nl/pub/51537

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promotor: Prof. R. Dekker, EPS-2013-293-LIS, http://repub.eur.nl/pub/41513

Staadt, J.L., *Leading Public Housing Organisation in a Problematic Situation: A Critical Soft Systems Methodology Approach*, Promotor: Prof. S.J. Magala, EPS-2014-308- ORG, http://repub.eur.nl/pub/50712

Stallen, M., *Social Context Effects on Decision-Making: A Neurobiological Approach*, Promotor: Prof. A. Smidts, EPS-2013-285-MKT, http://repub.eur.nl/pub/39931

Szatmari, B., *We are (all) the champions: The effect of status in the implementation of innovations*, Promotors: Prof J.C.M & Dr D. Deichmann, EPS-2016-401-LIS, http://repub.eur.nl/pub/94633

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promotors: Prof. D.L. van Knippenberg & Prof. P.J.F. Groenen, EPS-2013-280-ORG, http://repub.eur.nl/pub/39130

Tuijl, E. van, *Upgrading across Organisational and Geographical Configurations*, Promotor: Prof. L. van den Berg, EPS-2015-349-S&E, http://repub.eur.nl/pub/78224

Tuncdogan, A., *Decision Making and Behavioral Strategy: The Role of Regulatory Focus in Corporate Innovation Processes*, Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda, & Prof. T.J.M. Mom, EPS-2014-334-S&E, http://repub.eur.nl/pub/76978

Uijl, S. den, *The Emergence of De-facto Standards*, Promotor: Prof. K. Blind, EPS-2014-328-LIS, http://repub.eur.nl/pub/77382

Vagias, D., *Liquidity, Investors and International Capital Markets*, Promotor: Prof. M.A. van Dijk, EPS-2013-294-F&A, http://repub.eur.nl/pub/41511

Valogianni, K., *Sustainable Electric Vehicle Management using Coordinated Machine Learning*, Promotors: Prof. H.W.G.M. van Heck & Prof. W. Ketter, EPS-2016-387-LIS, http://repub.eur.nl/pub/93018

Veelenturf, L.P., *Disruption Management in Passenger Railways: Models for Timetable, Rolling Stock and Crew Rescheduling*, Promotor: Prof. L.G. Kroon, EPS-2014-327-LIS, http://repub.eur.nl/pub/77155

Venus, M., *Demystifying Visionary Leadership: In search of the essence of effective vision communication*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-289- ORG, http://repub.eur.nl/pub/40079

Vermeer, W., *Propagation in Networks:The impact of information processing at the actor level on system-wide propagation dynamics*, Promotor: Prof. P.H.M.Vervest, EPS-2015-373-LIS, http://repub.eur.nl/pub/79325

Versluis, I., *Prevention of the Portion Size Effect*, Promotors: Prof. Ph.H.B.F. Franses & Dr E.K. Papies, EPS-2016-382-MKT, http://repub.eur.nl/pub/79880

Vishwanathan, P., *Governing for Stakeholders: How Organizations May Create or Destroy Value for their Stakeholders*, Promotors: Prof. J. van Oosterhout & Prof. L.C.P.M. Meijs, EPS-2016-377-ORG, http://repub.eur.nl/pub/93016

Visser, V.A., *Leader Affect and Leadership Effectiveness: How leader affective displays influence follower outcomes*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-286-ORG, http://repub.eur.nl/pub/40076

Vries, J. de, *Behavioral Operations in Logistics*, Promotors: Prof. M.B.M de Koster & Prof. D.A. Stam, EPS-2015-374-LIS, http://repub.eur.nl/pub/79705

Wagenaar, J.C., *Practice Oriented Algorithmic Disruption Management in Passenger Railways*, Promotors: Prof. L.G. Kroon & Prof. A.P.M. Wagelmans, EPS-2016-390-LIS, http://repub.eur.nl/pub/93177

Wang, P., *Innovations, status, and networks*, Promotors: Prof. J.J.P. Jansen & Dr V.J.A. van de Vrande, EPS-2016-381-S&E, http://repub.eur.nl/pub/93176

Wang, T., *Essays in Banking and Corporate Finance*, Promotors: Prof. L. Norden & Prof. P.G.J. Roosenboom, EPS-2015-352-F&A, http://repub.eur.nl/pub/78301

Wang, Y., *Corporate Reputation Management: Reaching Out to Financial Stakeholders*, Promotor: Prof. C.B.M. van Riel, EPS-2013-271-ORG, http://repub.eur.nl/pub/38675

Weenen, T.C., *On the Origin and Development of the Medical Nutrition Industry*, Promotors: Prof. H.R. Commandeur & Prof. H.J.H.M. Claassen, EPS-2014-309-S&E, http://repub.eur.nl/pub/51134

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value*, Promotor: Prof. A. de Jong, EPS-2013-277-F&A, http://repub.eur.nl/pub/39127

Yang, S., *Information Aggregation Efficiency of Prediction Markets*, Promotor: Prof. H.W.G.M. van Heck, EPS-2014-323-LIS, http://repub.eur.nl/pub/77184

Ypsilantis, P., *The Design, Planning and Execution of Sustainable Intermodal Port-hinterland Transport Networks*, Promotors: Prof. R.A. Zuidwijk & Prof. L.G. Kroon, EPS-2016-395-LIS, http://repub.eur.nl/pub/94375

Yuferova, D., *Price Discovery, Liquidity Provision, and Low-Latency Trading*, Promotors: Prof. M.A. van Dijk & Dr D.G.J. Bongaerts, EPS-2016-379-F&A, http://repub.eur.nl/pub/93017

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promotor: Prof. M.B.M. de Koster, EPS-2013-276-LIS, http://repub.eur.nl/pub/38766

Zuber, F.B., *Looking at the Others: Studies on (un)ethical behavior and social relationships in organizations*, Promotor: Prof. S.P. Kaptein, EPS-2016-394-ORG, http://repub.eur.nl/pub/94388

Passengers, Crowding and Complexity was written as part of the Complexity in Public Transport (ComPuTr) project funded by the Netherlands Organisation for Scientific Research (NWO). This thesis studies in three parts how microscopic data can be used in models that have the potential to improve utilization, while preventing excess crowding.

In the first part, the emergence of crowding caused by interactions between the behavior of passengers and the public transport operators who plan the vehicle capacities is modeled. Using simulations the impact of the information disclosed to the passengers by public transport operators on the utilization and passenger satisfaction is analyzed. A quasi-experiment with a large group of students in a similar setting finds that four types of behavior can be observed.

In the second part, algorithms that can extract temporal and spatial patterns from smart card data are developed and a first step to use such patterns in an agent based simulation is made. Furthermore, a way to generate synthetic smart card data is proposed. This is useful for the empirical validation of algorithms that analyze such data.

In the third and final part it is considered how individual decision strategies can be developed in situations where there exists uncertainty about the availability and quality of travel options. We investigate how the best strategy for a specific type of objective can be computed. Finally, we analyze which strategies are worthwhile to consider for a very broad set of objectives.

ERIM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERIM

ERIM PhD Series
Research in Management

Erasmus University Rotterdam (EUR)
Erasmus Research Institute of Management
Mandeville (T) Building
Burgemeester Oudlaan 50
3062 PA Rotterdam, The Netherlands

P.O. Box 1738
3000 DR Rotterdam, The Netherlands
T +31 10 408 1182
E info@erim.eur.nl
W www.erim.eur.nl